Advancements in High-Throughput Mass Spectrometric Analyses of Complex Mixtures

by

Michael Joseph John Recchia

M.Sc. (Chemistry), UNC Wilmington, 2018

B.Sc. (Biology), UNC Wilmington, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Chemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2024

Declaration of Committee

Name:	**Michael Joseph John Recchia**
Degree:	**Doctor of Philosophy (Chemistry)**
Title:	**Advancements in High-Throughput Mass Spectrometric Analyses of Complex Mixtures**
Committee:	**Chair: Neil Branda** Professor, Chemistry
	Roger G. Linington Supervisor Professor, Chemistry
	David J. Vocadlo Committee Member Professor, Chemistry
	Andrew J. Bennet Committee Member Professor, Chemistry
	Vance E. Williams Examiner Professor, Chemistry
	Daniel Raftery External Examiner Professor, Anesthesiology and Pain Medicine University of Washington

Abstract

Recent advancements in high-throughput (HT) multi-omics have significantly impacted drug discovery, enhancing the exploration of novel chemistries through innovative analytical techniques. Metabolomics, a fundamental branch within the 'Omics' disciplines, focuses on the comprehensive and quantitative characterization of complex mixtures, often using mass spectrometry (MS) for analysis. Despite progress in multi-omics technologies, the chemical analysis of natural products (NP) mixtures remains a rate-limiting step in the discovery pipeline, creating a gap between chemical and biological profiling. Liquid chromatography (LC) is essential for high-quality chemical analysis of NP mixtures but can be time consuming for large sample libraries. Addressing the limitation, a dual-grid orthogonal sample pooling strategy called MultiplexMS was developed to increase LC-MS throughput. In a proof-of-concept study involving 925 fractionated NP extracts, MultiplexMS demonstrated efficacy by rediscovering all previously reported bioactive metabolites in only 5% of the original MS time. MultiplexMS-Q, an addition to MultiplexMS, introduces mathematical formulations to calculate relative quantitation data from MS features detected in pooled samples. This quantitative information enriches the qualitative data provided by MultiplexMS, offering a comprehensive view of NP mixtures. Advanced MS instrumentation can address sensitivity and mass accuracy limitations associated with chemically profiling complex mixtures. A subsequent study evaluated the benefits of high-resolution (HR) mass spectrometers, emphasizing improved resolving power, scanning speed, and sensitivity while maintaining accurate identification in complex samples. Integrating various omics techniques is crucial for modern NP discovery. Using a combination of metabolomics and genomics, new molecules with unique chemical scaffolds were discovered from under expressed biosynthetic machinery. These interdisciplinary approaches deepen our understanding of NP chemistry and accelerate drug discovery. MultiplexMS and MultiplexMS-Q represent advancements in HT chemical analysis, bridging the gap between biological and chemical profiling. The continued integration of omics data is essential for discovering novel compounds with potential therapeutic applications. These advancements empower researchers to efficiently uncover new bioactive molecules, contributing to developing new therapeutics.

Keywords: Natural Products; Metabolomics; High-Throughput; Mass Spectrometry; Liquid Chromatography; Deconvolution; Sample Pooling

This thesis is dedicated to my beloved wife, Alexis.

Without her unwavering support, this endeavor would not have been possible.

I love you.

Acknowledgements

First and foremost, I would like to thank Prof. Roger G. Linington for his exceptional guidance and mentorship throughout my Ph.D. His wisdom and genuine scientific curiosity have instilled in me a deep sense of inquiry into the natural world and the confidence to plan and execute solutions to problems in the laboratory and in life. I greatly admire your enthusiasm and resolve and appreciate your ability to turn potential setbacks into promising outcomes. My time during my master's degree taught me how to function in a laboratory, but under your tutelage, I truly learned to be a scientist. It has been an honor to study under your guidance and to work in such a nurturing environment to achieve my scientific ambitions.

I am very grateful to my committee members, Dr. David Vocadlo and Dr. Andrew Bennet, for their invaluable expertise and guidance.

Thank you to David Heywood, Adam King, Dr. Lee Gethings, and many others at Waters Corp. for their continued support of the Linington Lab, invaluable learning experiences, and opportunities to work with cutting-edge technologies. Visiting the Waters headquarters will forever be a highlight of my academic journey.

I am forever grateful for the opportunity to work with such amazing colleagues and friends in the Linington Lab: Dr. Trevor Clark, Dr. Sanghoon Lee, Dr. Joe Egan, Dr. Dennis Liu, Dr. Tim Baumeister, Claire Fergusson, Dr. Sandra Keerthisinghe, Liana Zaroubi, Brandon Showalter, Emily McMann, Hannah Cavanagh, Ella Poynton, Matthew Pin, Lisa Lin, and Julia Saulog. A very special thank you goes to Dr. Trevor Clark, Dr. Joe Egan, and Dr. Sanghoon Lee, who supported me at the start of my Ph.D. journey and made the lab environment enjoyable and exciting every day; to Dr. Tim Baumeister for his expertise in programming and knowledge about all things science; to Claire Fergusson for all the fun conversations and for keeping the lab running smoothly; to Dr. Dennis Liu for running endless bioassays and the engaging lab discussions; to Dr. Sandra Keerthisinghe for maintaining the HTCB facility; to Matthew Pin for your computer science expertise; to Liana Zaroubi and Brandon Showalter, fellow Ph.D. students who understand the struggle; to Lisa Lin for handling laboratory logistics; to Julia Saulog, without whom our lab would not run as efficiently; and to Emily McMann, Ella Poynton, and Hannah Cavanagh for your enthusiasm, constructive criticism, and energy. I thank you all and wish you the best in future research endeavors.

Table of Contents

List of Figures

List of Acronyms

ADE	Acoustic Droplet Ejection
AMI	Acoustic Mist Ionization
ATCC	American Type Culture Collection
BGC	Biosynthetic Gene Cluster
BHI	Brain Heart Infusion
BioMAP	Antibiotic Mode of Action Profile
CAM	Compound Activity Mapping
CLSI	Clinical and Laboratory Standards Institute
EIC	Extracted Ion Chromatogram
ESI	Electrospray Ionization
FDA	United States Food and Drug Administration
FPR	False Positive Rate
GT	Ground-Truth
HDAC	Histone Deacetylase
HPLC	High Pressure Liquid Chromatography
HR-MS	High-Resolution Mass Spectrometry
LB	Lysogeny Broth
LC-MS	Liquid Chromatography Mass Spectrometry
L-FDVA	Nα-(2,4-dinitro-5-fluorophenyl)-L-valinamide
MALDI	Matrix-Assisted Laser Desorption/Ionization
MIC	Minimum Inhibitory Concentration
MMS	MultiplexMS
MMSO	MultiplexMS Organization
MRT	Multi-Reflecting Time-of-flight
NCI	United States National Cancer Institute
NP	Natural Product
NPNPD	NCI Program for Natural Product Discovery
NRPS	Nonribosomal Peptide Synthetases
PK	Polyketide

PKS	Polyketide Synthetase
PUFA	Polyunsaturated Fatty Acid
qTOF	Quadrupole Time-of-flight
RiPP	Ribosomally Synthesized and Post-translationally Modified Peptide
RP	Reversed Phase
SPE	Solid Phase Extraction
TSP	Tryptic Soy Broth
UPLC	Ultra Performance Liquid Chromatography

Chapter 1.

Introduction

1.1. High-Throughput Mass Spectrometry

1.1.1. Natural Products Chemistry Discovery Pipeline

Over the past decades, massive advances in the drug discovery pipeline have revolutionized traditional methods of identifying therapeutic targets.[1–4] Improvements in chromatographic technology and the introduction of highly sensitive mass spectrometry (MS) instrumentation have enabled analysts to surpass previous analytical limitations.[5] The drug discovery and development pipeline is an interdisciplinary process encompassing multiple research phases to facilitate the creation of effective therapies. This process typically comprises target-based or phenotypic screening for hit identification and chemical characterization using spectroscopic techniques such as MS and nuclear magnetic resonance (NMR).[6–8] Subsequent chemical modifications optimize the activity and physicochemical profile of the target molecule, influencing the *in vivo* behavior, including potency, clearance, and safety.[3,9] However, discovering a potential target molecule involves extensive analysis and effort, often incorporating improved data analytics and predictive methods to associate chemical properties with biological activity. Researchers are increasingly focused on enhancing data handling and interpretation to create more efficient and accurate predictive models for drug discovery.[1,10–16]

Modern applications of the drug discovery pipeline often employ high-throughput screening (HTS) campaigns with large synthetic molecule libraries containing several hundred thousand compounds.[17] In natural products (NP) chemistry, researchers frequently work with prefractionated samples from crude environmental extracts, such as those derived from bacteria, fungi, and plants. The prefractionated extracts, which form large screening libraries and the precursor samples for further downstream analysis and fractionation steps, are often complex mixtures containing numerous primary and secondary metabolites. Prefraction and fraction refer to a daughter sample from a parent extract and the two terms are used interchangeably in this thesis. These libraries can be extensive, consisting of tens to hundreds of thousands of prefractionated samples,

typically separated by a chromatographic technique (**Figure 1.1**).[18,19] These prefractionated samples can potentially contain tens to hundreds of molecules in a single sample, equating to tens of thousands of MS features (detected ions with a specific mass-to-charge ratio, *m/z*) following MS analysis of large libraries. The preference for such prefractionated samples is the increased potential to identify bioactivity, which would likely be missed in crude extract screening due to the sample complexity and interference from other components in the mixture. For this purpose, identifying bioactive NPs and using their target engagement is often preferred from a prefractionated library rather than an NP crude extract.[18]

Although NP-based HTS campaigns have become less favored by biopharmaceutical companies, academic and government institutions are actively promoting these efforts.[10] For example, an initiative led by the United States National Cancer Institute (NCI) aims to stimulate HTS efforts and accelerate NP drug discovery through the NCI Program for Natural Product Discovery (NPNPD).[18–20] This program has developed a publicly accessible HTS-amenable library containing 1,000,000 fractions from 125,000 marine, microbial, and plant extracts collected globally. Other fractionated libraries for NP discovery campaigns include the Nature Bank (202,983 fractions)[21] and the Bioinformatics Institute Singapore (120,000 fractions).[22]

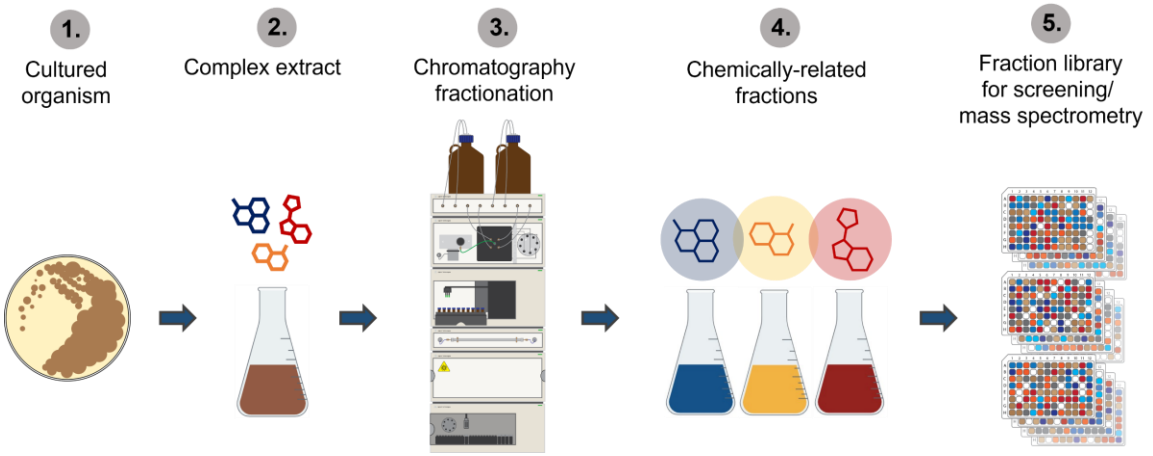


Figure 1.1 **Creating a Fractionated Library from a Biological Extract.**
The typical NP fractionation scheme illustrates creating content libraries from fractionated extracts of cultured organisms or collected biomass. (**1**) An organism (such as bacteria or fungi) is isolated from the environment and cultured. For plants or sponges, biomass is collected. (**2**) Production is scaled up to yield sufficient metabolite quantities for downstream processes, including metabolomics, structure elucidation, and bioassay screening. The metabolites are extracted from the large-scale production or physical biomass using organic solvents (e.g., MeOH, DCM). (**3**, **4**) The organic extract undergoes fractionation

using an aqueous organic elution gradient that separates compounds based on polarity and simplifies the samples for efficient mass spectrometry analysis. (**5**) Fractions are deposited into library plates for future bioactivity screening or metabolomics studies.

The early adoption of appropriate technologies in the drug discovery pipeline can significantly shorten the timeline from identifying potential therapeutic targets to pre-clinical development. Therefore, there is a concerted effort to develop and utilize advanced software and instrumentation to guide researchers toward effective targets, aiming to improve the success rate of identified therapeutic development.[6,23] This involves using novel sampling and investigative omics technologies to identify, target, and characterize potential candidates from complex mixtures of molecules.[16,24] Metabolomics, which focuses on the comprehensive and quantitative chemical characterization of complex mixtures, has particularly benefited from these advancements and has become essential in fields such as NP discovery.[25] However, advancements in high-throughput (HT) chemical characterization experiments of large complex sample libraries, particularly when using liquid chromatography (LC) before MS, have not kept pace with technological improvements in the omics field and, therefore, hinder the modern, fast-paced drug discovery efforts.

Chemical characterization of the extract samples is integral to the drug discovery process.[26] It is becoming increasingly apparent that NP discovery necessitates an integrative understanding of chemistry, biology, and computational methods. One-dimensional perspectives, such as bioassay-guided fractionation (BGF), are no longer sufficient to maintain the identification rate for novel bioactive scaffolds. New technologies and strategies (discussed later in **Section 1.1.4**) leverage a biological extract's chemical composition with the phenotypic profile to predict the active constituent(s).[27] Acquiring the complete metabolomic profile of a sample library can be advantageous for applying the sample set to various bioassays, reducing the time required to identify target compounds and assessing the distribution of a molecule across a sample library. However, herein lies the problem. Modest screening facilities can assay upwards of 50,000 samples a day in HT bioassays. Still, the chemical characterization of large extract libraries is a limiting factor, especially when applying LC-MS analyses, as the time needed for chromatography hampers the whole analysis pipeline.[28] For example, the problem with HT LC-MS analysis becomes apparent when deciding how to effectively sample the 1,000,000-member NPNPD sample library. Given

the traditional metabolomics workflow, samples should be analyzed as replicates to establish an accurate representation of each sample's chemical composition.[29] Applying this methodology to the entire NPNPD library, with an elution gradient of 5 minutes (8-minute complete cycle), it would take more than 15 years of continuous instrument use to obtain the full metabolomics dataset, a timeline not aligned with modern HT drug discovery.[6] Therefore, developing strategic approaches to mitigate time constraints in LC-MS analysis while preserving data quality is crucial for enabling a more efficient throughput capacity in the discovery pipeline. This objective forms the central thesis of this study.

1.1.2. Prospects in High-Throughput Natural Products Discovery

Interest in HTS and NPs for drug discovery remains an attractive research topic in academic settings despite waning interest in the pharmaceutical industry.[10,12] This lack of interest from pharmaceutical companies is mainly due to the complexities associated with NP screening, regulations on obtaining access to international bioresources, the high complexity of biological samples, limited success in HTS campaigns, restrictive NP patent laws, and the costly efforts needed to reduce this complexity to make the screening process more fruitful.[12,30] However, this has sparked enthusiasm among researchers to develop methods for more efficiently identifying chemical targets in extracts that could contain thousands of compounds. HTS campaigns are often applied in two aspects: biological screening, which investigates the impact of a molecule in a biological application against a disease target or mechanism, and chemical screening, which analyzes the chemical constitution of a sample.[26,31] Improvements in sensitivity and associated technologies, such as sophisticated instrumentation and laboratory automation, are rapidly enhancing the throughput potential of HTS applications.[32]

Phenotypic drug discovery has re-emerged as an effective strategy. It focuses on modulating a disease phenotype or biomarker rather than a pre-specified target to provide therapeutic benefits.[3] One notable application of this is Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes.[33–35] This technique is powerful for identifying small molecules as potential therapeutics and uncovering genetic regulators of various biological processes. Cell Painting allows for high-content image analysis of individual cells, measuring approximately 1,500 morphological features to produce a rich profile suitable for

detecting subtle phenotypes.[33,35] The advantage of image-based analyses is the ability to measure many samples and phenotypes simultaneously. These improvements in biological screening have significantly increased the throughput potential of analyses, outpacing other areas of the drug discovery pipeline. The challenge remains to bring the chemical characterization of complex extracts up to the same speed to realize the potential of HTS in NP drug discovery fully.

The second aspect of HTS is chemical characterization, which involves chemically annotating molecules in complex mixtures to determine the chemical constitution of samples. This is an application of metabolomics.[12,25] This process leverages analytical techniques, including MS and NMR spectroscopy, to analyze the chemical complexity of a sample, offering valuable insights into the structural characteristics of the molecules present. However, due to the insensitivity of NMR, MS is the most common analytical application in metabolomics, particularly in untargeted experiments, which aim to measure all the molecules in a sample unbiasedly.[12] Modern MS instrumentation analyzes complex mixtures with remarkable accuracy, providing high resolving power, sensitivity for detecting low-abundance molecules and ions, and rapid scanning speeds, enabling precise analysis of intricate sample compositions.[5,10] While direct infusion MS (DIMS) can be useful for targeted analyses, NP samples are often too complex, leading to compromised data quality. Therefore, a front-end application that simplifies sample complexity, like LC, is needed before MS analysis.[36] **Section 1.1.5** further discusses the advantages of using high-resolution (HR)-MS instrumentation to detect molecules precisely in complex mixtures and the analytical benefits of high-accuracy measurements in molecule dereplication.

1.1.3. Modern High-Throughput Methodologies

Due to its unparalleled precision, speed, and ability to identify complex mixtures, LC-MS is the most widely used method for untargeted metabolomics.[5] Modern advancements in LC capabilities offer high separation efficiency, improving downstream data analysis of complex samples. Ultra-performance liquid chromatography (UPLC) platforms, which use columns packed with sub-2 µm particles, require high pressures and provide improved separation performance over traditional LC methods.[5,37] While UPLC-MS offers exceptional quality analysis, it is important to be aware of its challenges. The time required to process samples chromatographically can be a

hindrance, especially with gradients as short as 5 minutes per sample. This becomes a bottleneck in drug discovery, particularly when biological screening can process similarly sized sample sets in a fraction of the time.[29]

Despite these challenges, improvements are underway in this area.[10,38] Analysts routinely implement ultra-high-throughput (uHT) MS analyses, a technology that leverages optimized ionization and sample delivery techniques to enable sampling from plates at a rate of over one sample per second.[1] These technologies, which do not require a chromatographic separation step, can be used in various screens to detect a broad range of analytes, including small molecules, lipids, and proteins. uHT techniques include acoustic droplet ejection (ADE)[39–41], acoustic mist (AM)[42], matrix-assisted laser desorption electrospray ionization (MALDI)[43,44], and nanostructure-initiator mass spectrometry (NIMS).[45] These methods improve sample throughput and allow screening libraries on a timeline comparable to biological screening campaigns. However, like the earlier arguments regarding DIMS, these techniques are best suited to targeted applications involving single-molecule or low-complexity samples.[11,44]

Another route for HT screening is affinity selection (AS)-MS.[46] This approach aims to separate binders from non-binders against a biological target. In AS experiments, compounds with high affinity are separated from those that do not bind to the target. Usually, each library member is present at minute concentrations (often pico- or femtomolar), with the target concentration driving the binding equilibrium. When an immobilized target is selected, the high local concentration on the surface can lead to a rebound effect, resulting in more efficient capture of binders.[46] This methodology integrates metabolomics analysis with biological screening, reducing the need to conduct bioassay screening and metabolomics analyses separately. However, while AS-MS is still developing and successfully applied in many HT applications, comprehensive MS analysis of all samples is limited by the lack of metabolite information in the screening application of AS-MS.[46]

The Agilent RapidFire HT-MS system is a sample delivery system that performs online solid-phase extraction (SPE) separation before MS analysis.[47] This SPE-MS process improves sample analysis throughput, removes salts and background components before analysis, and allows for the analysis of samples at approximately 9 seconds per sample, enabling the analysis of a 384-well plate in just over 1.5 hours.

Users of this technology can apply sampling with various techniques, including lead discovery, forensic toxicology, and multi-omic applications.[47–49] This flexibility is due to the interchangeability of the SPE cartridges with different stationary phases suitable for the study. The RapidFire application has been adapted to AS-MS applications to increase therapeutic discovery.[50] However, SPE-MS has limitations, especially in analyzing high-complexity NP samples, such as the lack of chromatographic separation, reduced ionization efficiency, limited quantitative accuracy, and less effective untargeted profiling compared to targeted analyses.[51] Nevertheless, these limitations do not diminish the RapidFire system's value as an effective tool for label-free HT drug discovery.

In UPLC-MS analyses, single-channel systems are commonly used, where a single column performs chromatography on each sample sequentially. Technological developments have aimed to increase the utilization of MS scanning time by parallelizing the chromatographic method.[52] This involves using multiple columns simultaneously in independent LC channels to accelerate the speed of analysis. This technique, known as multi-channeling, can increase throughput over single-column chromatography while retaining high chromatographic quality.[52] However, though this system's throughput is advantageous for smaller libraries, it is still insufficient for the extensive libraries discussed earlier. Additionally, these systems can be costly and complex to operate correctly.

Until these techniques can match the separation potential of UPLC-MS, chromatographic separation will continue to be the mainstay of metabolomics in large screening campaigns. UPLC-MS excels in analyzing samples requiring low detection limits from complex matrices or characterizing complex biotherapeutics such as antibody-drug conjugates.[6,12,36] The reduced chromatographic dispersion of eluting analytes promotes improved source ionization efficiency.[15] Large-scale HT experiments drive biotechnology research and the search for drug candidates, offering experimental systems that automate the discovery process and promptly allow for tens of thousands of samples to be analyzed. Improvements in instrumentation and software processing data have reduced the likelihood of experimental errors, providing efficient and robust data analysis. Setting aside a significant bottleneck that hinders the NP discovery pipeline, namely structure elucidation, the research described herein targets the

throughput capabilities of UPLC-MS analysis of large, complex sample libraries by reducing the time needed for chromatographic analysis while maintaining data quality.[12]

1.1.4. Metabolomics Software Development and Data Repositories

Modern software applications and data repositories greatly enhance high-throughput metabolomics analysis. These tools are crucial in accurately deconvoluting MS features in complex mixtures and aligning samples in experiments to map molecules across datasets. The emergence of data repositories and chemoinformatic approaches has significantly expedited the dereplication of NP discovery, enabling the distinction of known chemistry from novel compounds within a sample.

The Natural Product Atlas (NP Atlas), a globally curated database housing over 36,000 distinct microbial NP compounds, is an important asset in NP discovery. The database provides essential structural and chemical information based on published data to expedite discovery and dereplication processes.[53] NP Atlas also offers comprehensive analyses of reported compounds, providing links to additional databases such as MIBiG 3.0[54], SMART 2.0[55], and NP Classifier.[56] At the same time, tools such as CFM-ID predict MS/MS fragmentation spectra, and databases like METLIN offer empirical MS/MS data.[57–59] NMR-based metabolomics has also gained traction due to advancements in instrumentation and computational algorithms, supported by resources like the Natural Products Magnetic Resonance Database (NP-MRD), which archives NMR spectra of identified compounds and can aid in the development of computer-assisted structure elucidation (CASE) tools such as DP5.[13,60–62]

Data preprocessing platforms have significantly advanced due to innovations in deconvolution algorithms, hardware improvements, and reference datasets, which facilitate dereplication or association with known families or classes of compounds.[14] These platforms transform raw data from the MS acquisition into a machine-friendly format suitable for analysis by various bioinformatic software packages. In metabolomics, preprocessing steps include deconvolution, calibration, normalization, alignment, lock mass correction, and missing value imputation.[12,14] Many platforms exist for data preprocessing, performing the usual steps with great accuracy and precision and often tailored to specific instrument vendors and data acquisition methods. This step is crucial for aligning MS features between samples and observing the relative

abundances of molecules across a sample set. However, despite the precision of these platforms, processing time remains a limiting factor, especially when analyzing large datasets or thousands of samples.[14]

Modern bioinformatic approaches have surpassed traditional one-dimensional methods, enabling the rapid identification of novel bioactive scaffolds in complex mixtures. These approaches allow researchers to identify potential bioactive candidates more efficiently, reducing the need for iterative testing and isolation. For example, Kurita *et al.* combined NP chemistry, metabolomics, HTS, and similarity networking to discover a novel family of NPs from extracts saturated with known compounds.[63,64] NP Analyst, an open-source platform for Compound Activity Mapping (CAM), integrates metabolomic and bioactivity data to prioritize compounds.[27] The Global Natural Products Social Molecular Networking, or GNPS, provides an open-access database of community-shard MS/MS spectrometry data and molecular networking to visualize and annotate the chemical space in untargeted metabolomics.[57,65] McCaughey *et al.* developed the platform Isoanalyst that leverages stable isotopic labeling to correlate MS-based metabolomics data with bioinformatic predictions of compound structures.[66] In addition, various bioinformatic utilities complement this tool, including antiSMASH 6.0[67], an updated version of the most widely used tool for mining microbial genomes for secondary/specialized metabolite biosynthetic gene clusters (BGCs), and MIBiG 3.0[54], an online reference database of BGCs. These platforms, and others not mentioned here[68–70], underscore the critical need for rapid data acquisition in bioassay and UPLC-MS domains, highlighting the advanced software tools available to integrate HT capabilities. Software platforms have facilitated NP discovery, but there is still a gap in UPLC-MS sample throughput that must be addressed to leverage the processing and analysis capabilities of modern bioinformatic tools.

1.1.5. Mass Spectrometry Advances in Natural Products Mixtures

Advancements in various technologies highlight the capabilities of MS systems to analyze complex mixtures precisely. Over recent decades, MS has undergone remarkable advancements, revolutionizing the capability of investigating complex mixtures with unprecedented sensitivity, resolution, and throughput.[71] Since its inception in the early 20th century, MS has become the central technique extensively utilized for

molecular analyses.[72] It has demonstrated distinct capabilities in accurately and sensitively analyzing complex mixtures, finding wide applications across various studies.

MS technology has driven data acquisition, processing, and interpretation innovations through advanced computational tools and bioinformatics pipelines. Introducing ionization techniques like ESI and MALDI has enabled the direct analysis of biomolecules such as peptides, proteins, and metabolites from complex biological matrices.[5,73] This foundation has been crucial for modern proteomics and metabolomics. Simultaneously, the development of high-resolution (HR) mass analyzers—such as time-of-flight (TOF), quadrupole-TOF (qTOF), ion trap, and Orbitrap analyzers—has expanded the capacity of MS to resolve complex mixtures with enhanced accuracy and mass resolving power.[74,75] Integration with chromatographic techniques such as LC further enhances the resolving power and analytical throughput of MS-based analyses, enabling comprehensive characterization of complex samples with diverse chemical compositions.[5] These advancements have been pivotal in overcoming the challenges posed by the complexity and diversity of biological samples, making MS indispensable in research analyses. MS technology finds widespread application in diverse scientific domains, notably in metabolomics, where it provides insights into metabolic pathways, biomarker discovery, and disease mechanisms, significantly advancing our understanding of metabolic networks under various conditions.[5,73] Moreover, the ability of MS instrumentation to detect and characterize trace analytes with high specificity has driven breakthroughs in biomedical research and environmental monitoring.[74]

Innovative additions to MS instrumentation have significantly enhanced sensitivity and resolution capabilities. Techniques like ion mobility spectrometry (IMS) and ambient ionization methods continue to expand MS utility in real-time analysis of complex samples, bridging the gap between analytical chemistry and practical applications.[76] Modern instruments, exemplified by the Waters™ Multi-Reflecting Time-of-Flight (MRT), achieve remarkable precision (parts per billion (ppb) mass accuracy) and resolution (300,000 full width at half maximum (FWHM), enabling comprehensive analysis of complex mixtures and facilitating crucial processes such as dereplication and molecular formula determination.[77]

The evolution of MS technology has profoundly impacted scientific research, enabling discoveries across disciplines. As MS continues to evolve and integrate with

interdisciplinary approaches, it promises to uncover novel insights into biological systems, environmental dynamics, and material sciences, thereby shaping the future of scientific inquiry and innovation. **Chapter 4** discusses using state-of-the-art MS technology to analyze increasingly complex NP mixtures and compares results to a current industry-leading HR mass spectrometer.

1.2. Advancements in High-Throughput Sampling

1.2.1. Introduction to Sample Pooling

The standard approach taken by HT experiments is to test each item individually in single analyses.[6] This approach, while thorough, is time-consuming and costly, mainly when the goal is to identify a relatively small number of items of interest from an extensive sample library. This issue is prevalent across various fields, including biological screening and public health screening, where timely analysis is crucial to, for instance, prevent the spread of a disease agent.[78,79] One proposed solution is sample pooling or multiplexing to address the challenge of low throughput.[80] Scientific implementation of the sample pooling approach was first described during World War II, when draftees were tested for syphilis.[80,81] In the approach, blood samples from multiple cadets were pooled and tested as one sample for the syphilitic antigen. If the test was negative, the cadets were cleared. Otherwise, a positive result would require retesting of individuals for the antigen. (**Figure 1.2**).[82,83] While pooling multiple samples and analyzing them simultaneously is generally effective, it also presents practical challenges. These challenges necessitate innovative experimental designs to minimize the need for reanalyzing samples and ensure accuracy during pooled sample analysis.[84] This opens opportunities to identify combinations or organizational strategies that enhance positive outcomes by reducing false positive rates (FPRs) or avoiding omitting positive results. Several applications of sample pooling for HT operations have been documented. However, there are currently no known applications specific to HT sample pooling in metabolomics analysis that efficiently handle large sample libraries using UPLC-MS.[85–88]

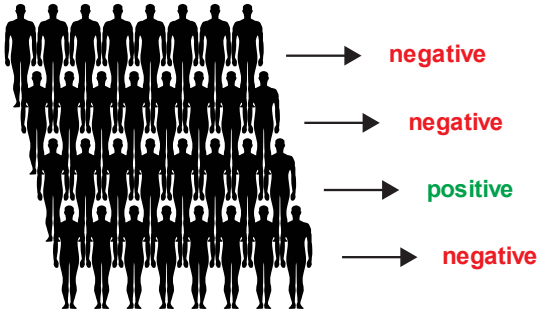


Figure 1.2 Example of Sample Pooling to Increase Throughput
In this sample pooling example, specimens from human subjects are combined into groups of eight for testing in a targeted assay. If the pooled sample tests negative, all individuals are considered negative. However, if the pooled sample tests positive, each subject within the pool is retested individually. This pooling approach has been applied in medical assays to streamline testing for large populations, reducing the required time and resources.[78,79]

1.2.2. Pooling Constraints in Untargeted Metabolomics

Throughput presents a notable constraint in metabolomics analyses. When employing untargeted metabolomics with UPLC-MS on biological specimens, each sample can yield thousands of individual signals known as "MS features," characterized by distinct *m/z* values and retention times.[10] Identifying which MS features represent the same analyte across multiple UPLC-MS runs becomes challenging, especially when profiling numerous samples simultaneously. Constructing MS feature lists for each sample is critical in untargeted metabolomics analysis. However, pooling multiple samples into a single analysis complicates the reconstruction of individual sample lists since the origin of each MS feature becomes ambiguous. This necessitates the development of characteristic sampling strategies to enable multiplexing and the identification of the origin of specific MS features for downstream analysis.

Sample pooling in untargeted UPLC-MS metabolomics presents challenges, but it is a strategy that can be carefully considered for specific applications. For instance, it is most suitable for applications where positive results, or the presence of a particular molecule, are infrequent in a dataset. This significantly reduces false positives upon the reconstruction of individual sample lists. HT NP experiments for discovering bioactive compounds align well with this strategy, given the typically low hit rates (<1 %) in biological NP screens.[12]

One multiplexing strategy addressing throughput limitations in HT metabolomics involves using pooled quality control (pooled QC) samples to establish a combined peak list of all analytes in a sample set.[89] This strategy involves combining small aliquots from individual samples to create a composite sample, which is then analyzed using UPLC-MS amongst the rest of the sample set. The pooled QC approach offers several advantages: it reduces sample variability by averaging individual differences, leading to more robust and reproducible data; it conserves resources by reducing the number of analytical runs needed, which is particularly beneficial in large-scale studies with limited sample volumes; and pooled samples can serve as quality control references throughout the analytical process, aiding in monitoring instrument performance and data consistency.[90] For instance, Stancliffe *et al.* recently proposed a data acquisition strategy capable of analyzing thousands of samples using GT peak lists from a pooled QC sample to identify features from batches of research samples.[89] While the pooled QC strategy is effective in many ways, it is important to note that it does have limitations. For instance, pooled QC eliminates the need for technical replicates for all samples, but the number of MS runs increases linearly with the sample count. This can be a drawback when meeting the needs of HT-MS screening. However, this pooling design is particularly effective for primary metabolomics applications where samples (e.g., urine or plasma) are inherently similar but vary in their analyte concentrations. In such cases, the method proves effective because the composition of the pooled QC sample mirrors that of individual samples, with only the relative abundances of analytes necessitating individual runs.

While throughput remains a significant constraint in untargeted metabolomics analyses, strategic sample pooling can help mitigate some challenges. Although pooled QC offers data robustness and resource efficiency, it is limited in resolving the increased MS run demands in HT-MS screening campaigns. Exploring advanced sample pooling strategies holds promise for improving throughput accuracy and efficiency in metabolomics analyses.

1.2.3. Orthogonal or Self-Deconvoluting Matrix Strategy

HT-MS screening campaigns often use a single-sample approach, which, while straightforward, becomes inefficient with large libraries where traditionally only a few active compounds are present.[86,91] Sample multiplexing strategies have been explored

to address this inefficiency, with varying success depending on the pooling scheme and application. One effective strategy for HT screening is orthogonal pooling, also known as the self-deconvoluting matrix (SDM) strategy.[84,87] Orthogonal pooling involves testing each sample twice as a component within different sets of samples. For an *n*-sample library tested in pools of size *r*, this strategy reduces the number of samples to be analyzed to $\frac{2n}{r}$. This represents an *r*-fold improvement in sampling efficiency compared to single-sample HT experiments using *n* samples with quadruplicate analysis.

The orthogonal pooling strategy uses a symmetrical matrix of *r* x *r*, where samples are assigned theoretical wells within the matrix.[84] These assigned samples are then pooled across each row and column. If a positive result is detected in both a row and a column pool, the positive designation is assigned to the sample at the intersection of that row and column. This methodology can be readily applied to NP mixtures, allowing detected molecules in row-column combinations to be assigned to the sample at the intersection of the pair (**Figure 1.3**). While there is no physical limit to the number of samples that can be pooled together, the complexity of samples and the potential for poor data quality during UPLC-MS must be considered. This complexity can be mitigated by strategically assembling samples within the grid to maximize the diversity of pooled samples, using HR instrumentation, and using complementary IMS during analysis.[37]

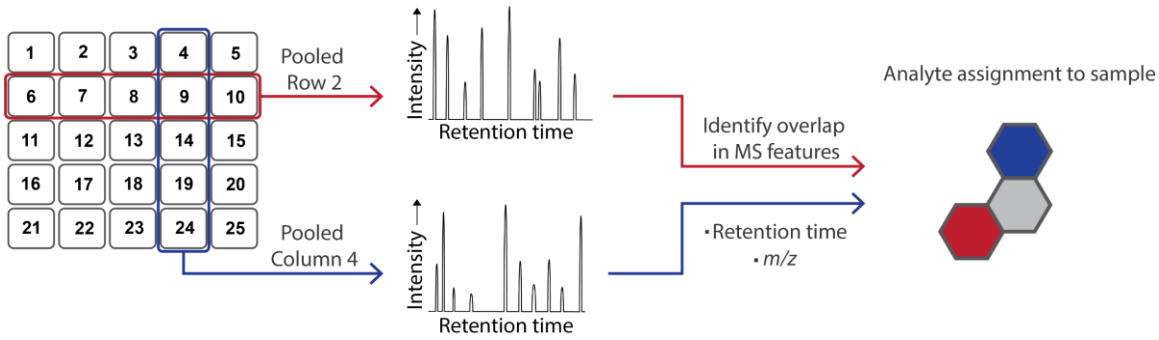


Figure 1.3 **Demultiplexing Strategy Using the Self-Deconvoluting Matrix.** Orthogonal sample pooling is a powerful technique for sampling complex NP mixtures. In this example, 25 samples are arranged in a 5 x 5 matrix. The rows and columns are independently pooled and analyzed by UPLC-MS. If a molecule is detected in a row and column, then the molecule is assigned to the sample at the intersection of the pair. This approach optimizes sampling efficiency and facilitates NP analysis.

Orthogonal pooling is particularly advantageous for NP screening, where libraries of fractionated extracts from different sources contain a vast array of structurally diverse compounds.[86,91] This HT sampling strategy is ideal for identifying rare molecules sparsely distributed across large extract libraries, as the FPR of scarce molecules is low compared to molecules with a high presence in a dataset. With the growing sizes of modern extract and compound libraries, the scalability of orthogonal sample pooling allows the accommodation of large libraries, making it particularly suitable for extensive NP screening by UPLC-MS.

However, biological extracts, particularly from plants, can contain multiple molecules or classes of molecules with similar or identical physicochemical properties. Pooling these types of samples can increase the risk of false positives during the demultiplexing step, as molecules detected in multiple pools may be wrongly assigned to multiple sample locations within the grid. Additionally, NP extracts are reputable for sample complexity, containing hundreds to thousands of molecules.[6] The combination of these dense samples can impact data quality in UPLC-MS analyses. While this complexity is typically addressed by a primary fractionation step before MS analysis, some fractions may contain more molecules than others. A possible multiplexing strategy in this instance involves pooling more complex fractions with less complex ones to improve data quality in HT experiments. For example, strategically pooling polar fractions with less polar ones reduces the chromatographic overlap of metabolites. Another effective strategy involves varying source samples to maximize metabolite diversity and minimize redundancy in pools. The rich diversity of NP samples, especially from bacteria, fungi, and sponges, makes orthogonal pooling ideal for uncovering rare molecules sparsely distributed in an extensive extract library.

Orthogonal pooling, or the SDM strategy, significantly improves throughput over traditional single-sample HT-MS screening methods.[84] This method enhances efficiency and accuracy in identifying rare compounds within large libraries by reducing the number of wells and leveraging a symmetrical matrix approach.[86] Strategic implementation and careful sample selection can mitigate pool complexity and reduce the FPR. However, further enhancement of the orthogonal strategy is needed to improve the accuracy of matrix deconvolution. This strategy shows that sample pooling is a promising method to enhance UPLC-MS sample throughput when adapted for large NP extract libraries.

1.2.4. Dual-Grid Orthogonal Sampling Pooling

Orthogonal sample pooling offers several advantages, but using a single matrix and pooling complex mixtures can complicate the demultiplexing process for detected molecules and sample assignments. A complementary strategy known as dual-grid orthogonal sample pooling can be implemented to address this. This method uses a second symmetrical grid as a counterpart to the initial grid.[84] Samples from the initial grid are then rearranged so that no two samples fall in the same row or column within the rearranged grid. While this strategy doubles the number of pooled samples ($\frac{4n}{r}$), the method increases the chemical diversity of pooled samples and provides intrinsic quadruplicate analysis of samples.[84] If a molecule is detected in two row and column pairs (totaling four pooled samples), and the sample identification is the same at both row and column intersections, the molecule can be confidently assigned to that sample.

Dual-grid orthogonal pooling reduces the required runs, enabling faster data acquisition and preprocessing of UPLC-MS analysis. Fewer runs mean less consumption of reagents, solvents, and instrument time, resulting in cost savings and a smaller environmental footprint. This approach is particularly advantageous for resource-limited labs, allowing more extensive research within budget constraints. However, successful implementation requires meticulous planning and data management. Each sample must be accurately tracked through its row and column pools to ensure precise deconvolution. Accurate mixing and precise handling techniques are essential to prevent cross-contamination and ensure representative sampling. Tailoring the pooling method may become necessary, considering the specific characteristics of the samples under analysis—such as solubility, stability, and their interaction with the UPLC-MS system.[6] Leveraging modern laboratory automation and customized data-handling scripts can reduce the risks associated with these factors and enhance the screening capabilities of HT campaigns.

The dual-grid orthogonal sample pooling method represents a promising strategy to advance NP UPLC-MS analysis. This strategy addresses key challenges in analyzing complex NP mixtures by enhancing throughput, improving data quality, and reducing costs. As technology evolves, the dual-grid methodology may become integral to HT NP research, facilitating the discovery of novel bioactive compounds with greater efficiency and precision. This thesis explores the practical applications, advantages,

disadvantages, and limitations of the dual-grid orthogonal sampling strategy in profiling large NP sample libraries and re-identifying bioactive molecules.

1.3. Conclusion

HT-MS has revolutionized the NP drug discovery pipeline, significantly improving the efficiency and effectiveness of identifying potential therapeutic compounds.[6,10] Integrating advanced chromatographic techniques with highly sensitive MS instrumentation has overcome previous analytical limitations, enabling rapid screening and characterization of large, complex sample libraries. This has been particularly transformative in NP chemistry, where researchers often deal with intricate mixtures from environmental extracts.[6]

Despite these advancements, a key challenge remains in aligning the speed of chemical characterization with biological screening. Traditional methods such as UPLC-MS provide high-quality data but are limited by the time-consuming nature of chromatographic separation. This bottleneck becomes pronounced when analyzing extensive libraries, such as those created by the NPNPD, which contains over a million fractions from diverse biological sources.[20] To address these challenges, researchers are developing innovative strategies to enhance throughput without compromising data quality.[12,32] Two promising approaches are uHT-MS and dual-grid orthogonal sample pooling. uHT-MS methods bypass chromatographic separation and significantly increase sample throughput but are best suited for target analysis with samples containing only a few molecules.[38] The dual-grid orthogonal sampling strategy involves pooling samples in rows and columns from symmetrical matrices.[84] These matrices contain the same samples but are arranged in different configurations to avoid overlap of the samples between grids. The dual-grid approach effectively reduces the required analyses compared to single-sample approaches. Additionally, the pooling and subsequent deconvolution platform are well-suited for complex mixtures. This sample pooling approach is well-suited for researchers wanting to obtain metabolomics data on large libraries of samples, so the chemical constitution is known prior to bioactivity testing, especially when testing the library against multiple targets. In contrast, this methodology may not be applicable to sample sets containing only a few samples as the time saving benefit would not be as advantageous to individual MS analysis.

Integrating sophisticated software and data repositories has facilitated the accurate deconvolution and analysis of MS data.[14,36] Computational platforms and curated databases, such as the Natural Product Atlas and GNPS, expedite dereplication by identifying known compounds and aiding in discovering novel bioactive molecules.[53,57] These tools enhance the interpretive power of MS data, allowing researchers to rapidly associate chemical features with identities, facilitate the annotation of large libraries, and dereplicate known molecules.

Ongoing advancements in HT-MS technologies, innovative data handling, and analytical strategies are reshaping the NP drug discovery landscape. While challenges persist, continuous developments promise to streamline the pipeline, accelerating the identification and development of new therapeutic agents. The future of NP drug discovery will likely emphasize multidisciplinary approaches, leveraging technological innovations and computational tools to meet the growing demand for efficient and effective drug discovery processes.

This thesis investigates a novel multiplexing strategy for analyzing large extract libraries using UPLC-MS to profile complex NP samples chemically. The approach uses dual-grid orthogonal pooling to maximize diversity within sample sets and leverages high-powered MS instrumentation to improve the profiling coverage. The following chapters explore the multiplexing strategy's limitations and advantages, effective sample organization approaches, relative quantitation calculations from pooled samples, and the potential to discover biologically active metabolites in large extract libraries.

Chapter 2 describes the strategy, design, and application of MultiplexMS – a novel MS-based multiplexing approach for the uHT analysis of complex mixtures. This chapter outlines the underlying theory behind the multiplexing strategy and how samples are strategically organized to diversify the sample pooling. The HT pipeline is then applied to rapidly acquire qualitative information about the chemical complexity of a large NP library, offering valuable insights for examining extensive collections. Notably, the chapter also discusses the development of a user-friendly graphical user interface (GUI), making this platform accessible and inclusive, even to non-programmers.

Chapter 3 continues Chapter 2, entitled “MultiplexMS-Q: A Quantitative Mass Spectrometry Method for Relative Feature Abundance Determination in Multiplexed

Mixtures.” MultiplexMS-Q enhances the original MultiplexMS approach, which is limited to providing qualitative data by generating presence and absence values for detected features (1/0). The new utility, MultiplexMS-Q, determines relative quantitative data using matrix algebra and least squares algorithms. This advancement facilitates the identification and prioritization of bioactive compounds in NP discovery, offering improved accuracy following computational deconvolution and comprehensive analysis of complex mixtures. Details are provided regarding the development of the MultiplexMS-Q pipeline and the integration into the existing GUI package.

Chapter 4 explores the use of various HR-MS instruments to assess the impact of different degrees of MS instrumentation on the outcomes of MultiplexMS computational deconvolution. Specifically, attention centers on the precision of assignment and mass accuracy of detected molecules. Comparative testing occurs on two instruments: the Waters SYNAPT G2-Si and the Waters Xevo MRT. The Xevo MRT is a brand-new qTOF instrument with impressive features, including a 100,000 FWHM resolving power and a mass accuracy of less than 500 ppb.

In **Chapter 5**, the focus shifts to the use of various analytical methods and omics technologies. The diverse set of exploratory techniques aimed to predict, isolate, and structurally elucidate novel NPs from the genome of a terrestrial bacterium strain, *Burkholderia megapolitana* FERM BP-3421.

Chapter 2.

MultiplexMS: A Mass Spectrometry-Based Multiplexing Strategy for Ultra-High-Throughput Analysis of Complex Mixtures

Published manuscript

Reprinted with permission from *Analytical Chemistry* **2023** *95*(32), 11908 – 11917.

DOI: 10.1021/acs.analchem.3c00939.

Authors: Michael J. J. Recchia, Tim U. H. Baumeister, Dennis Y. Liu, Roger G. Linington

Author Contributions: M.J.J.R. and T.U.H.B. made equal contributions to this project and contributed to the design of the study alongside R.G.L. M.J.J.R. led the mass spectrometric analysis, conducted follow-up microbial cultures, and spectroscopically identified isolated molecules. T.U.H.B. developed the code for *in silico* studies and the MultiplexMS application. Additionally, D.Y.L. and T.U.H.B. created custom scripts for worklist generation, essential for use with automated liquid handlers. D.Y.L. also assisted with the automated liquid handling system. M.J.J.R., T.U.H.B., and R.G.L. jointly authored the manuscript, and all authors approved the final version.

2.1. Introduction

In recent years, high-throughput (HT) multi-omics techniques have changed how the field pursues the discovery of natural products (NP). The methods are vital for prioritizing and directing the isolation of new chemical compounds from biological sources.[92,93] Among these techniques, metabolomics, an omics branch focused on the comprehensive and quantitative chemical characterization of complex NP mixtures, has thrived due to technological advancements and remains at the forefront of NP research.[94] Untargeted metabolomics analysis is fundamental for NP profiling as it leverages the exceptional sensitivity of mass spectrometry (MS) to reveal chemical information on large numbers of metabolites in biological samples.[25] However, despite its application potential, acquiring MS for large extract libraries remains a bottleneck in

discovery programs. The challenge lies in the time-intensive chromatographic separation required for complex mixtures. Despite the capacity of academic high-throughput screening (HTS) facilities to analyze over 50,000 samples daily, limitations in MS data acquisition persist. The analysis of such a sample library using MS would occupy a mass spectrometer for months of continuous use.[28] Consider the ambitious goal of the National Cancer Institute's (NCI) Program for Natural Products Discovery, which aims to build a 1,000,000-member fraction library. Even with a minimal elution gradient (as short as 5 minutes), ultra-performance liquid chromatography (UPLC)-MS analysis of this library would demand more than nine years of continuous instrument time. This estimate does not account for replicate analyses, which would further extend the processing duration.[20,29] While HT multi-omics techniques have revolutionized NP discovery, addressing the time-intensive nature of MS data aquation remains essential for accelerating the identification of valuable chemical entities.

Recent advancements in high-speed sampling techniques, including matrix-assisted laser desorption/ionization (MALDI),[44] acoustic mist ionization (AMI),[95] and acoustic droplet ejection (ADE),[40] have revolutionized sample delivery for rapid analyses (typically <10 s). These techniques excel in targeted analysis of samples with few analytes per sample, outperforming traditional liquid chromatography (LC) methods in specific experiments.[44] However, UPLC-MS methods offer improved coverage for complex samples like NP extracts, thanks to the additional dimension provided by chromatographic separation.

In public health screening, sample pooling is one practical solution to the challenge of low sample throughput. In this approach, tests are conducted on pools of samples, and individual samples are retested only if the pool returns a positive result.[78,79] This strategy is particularly suitable when the frequency of positive results is low, as screening multiple negative pools significantly reduces the number of tests required.[86] The discovery of bioactive NPs aligns well with this approach. Biological NP screens typically yield hit rates below 1%, encompassing diverse classes of NP structures. Given that individual NP structures are sparsely distributed across extract libraries, a pool/deconvolute approach is appropriate for compositional analysis. Surprisingly, despite the long history of sample pooling applications in other fields, no MS-based method for sample pooling has been developed for NP extract libraries.[80,82,85,86]

This study introduces MultiplexMS (MMS), a novel dual-grid orthogonal multiplexing strategy. The method addresses the HT limitations encountered in untargeted MS analyses by leveraging highly sensitive MS instrumentation through a sample pooling approach. Inspired by previously described methods, the multiplexing strategy efficiently analyzes NP extracts by pooling rows and columns from sample grids. Each pooled sample, comprised of multiple samples, is subjected to UPLC-MS. A subsequent computational workflow deconvolutes the resulting MS data from the pooled samples, reconstructing feature lists for each sample within the pool (**Figure 2.1**).[78,80,84,86] The MMS workflow capitalizes on the presence or absence of an MS feature (encoded as a 1 = present, 0 = absent), forgoing quantitative information to enhance throughput efficiency. This technique proves ideal for untargeted NP metabolomics, especially when identifying rare molecules within complex mixtures. However, it is unsuitable for quantitative or targeted metabolomics, which requires accurate concentration determinations across a broader sample range.[96,97] MMS enables large-scale MS analysis of extract libraries in a single experiment by pooling multiple samples, offering scalability and efficiency compared to traditional UPLC-MS methods. Complementing the process is a dedicated stand-alone software package that manages pool design and MS feature deconvolution. The study provides detailed insights into the development of MMS, covering proof-of-concept validation, exploration of pool size limits, *in silico* modeling to enhance pool design, creation of an open-source software platform for MMS, and validation through the rediscovery of bioactive compounds from a 925-member extract library.

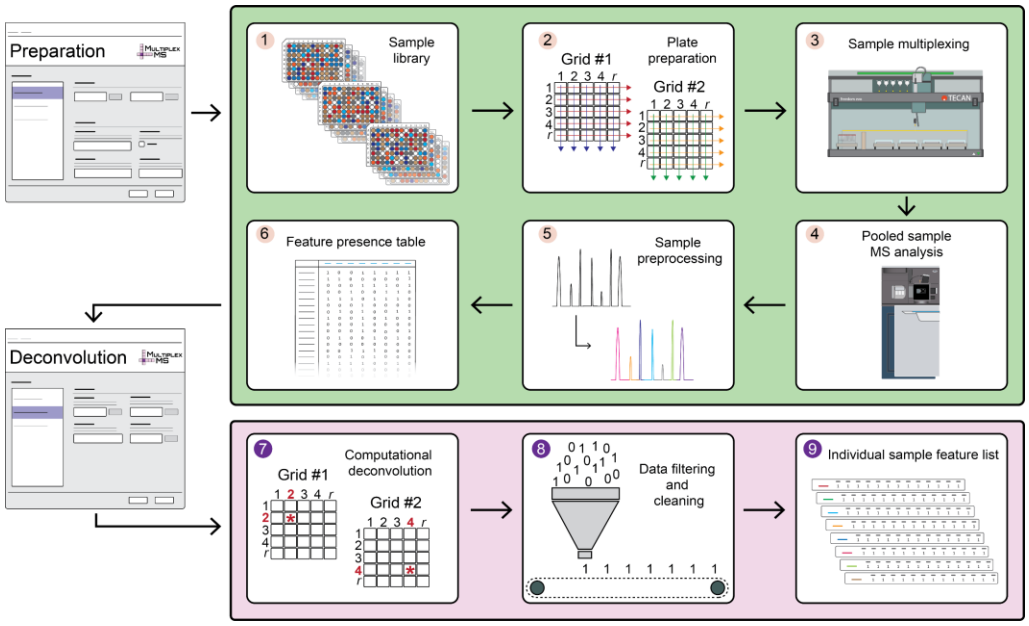


Figure 2.1 **Experimental Schematic of the MultiplexMS Screening Pipeline.** (**1**) Lists of samples from a fractionated extract library are provided to the MultiplexMS application, (**2**) Preparation grids and tables are generated using the 'Preparation' table in the MultiplexMS application, (**3**) Samples are pooled either manually or using automated liquid handling robotics, (**4**) Multiplexed samples are analyzed by UPLC-MS, (**5**) The MS data is preprocessed using a software platform that performs lock mass correction, spectral alignment, and peak picking, (**6**) An MS feature list is provided to the MultiplexMS application, (**7**) Computational deconvolution of pooled samples, (**8**) Optional data filtering and cleaning, (**9**) Feature lists for each sample in the grid are generated.

2.2. Results

2.2.1. Development of the Multiplexed Sampling Strategy

MMS revolutionizes UPLC-MS sample analysis by employing a novel dual-grid pooling strategy tailored for maximum throughput without compromising MS data integrity (**Figure 2.2**). Inspired by array testing methods[78], this approach organizes an *n* sample library into two MultiplexMS Organization (MMSO) grids, each configured as *r* x *r*, representing the initial and rearranged layouts (**Figure 2.3**). Samples are grouped into pools of size *r* and strategically positioned to ensure no overlap in row/column placement between the two grids. Using the dual-grid strategy minimizes the inherent challenges in deconvoluting identical MS features across multiple samples within a single grid. This strategy mitigates the risk of false positive assignments and inflated feature lists during sample reconstruction. For example, when multiple samples share the same MS feature, the deconvolution process can be prone to ambiguous interpretation, leading to

inaccurate feature assignments (**Appendix Figure A.1**). Despite doubling the number of samples for analysis, the incorporation of the second grid enhances the precision of MS feature assignment and reduces the false positive rate (FPR) during computational deconvolution (**Appendix Figure A.2**). The pooling process involves combining rows and columns from the initial and rearranged grids into pools of size *r*, where each sample makes up 1/*r* of the pooled samples. This ensures consistent sample concentration across analyses, preventing dilution effects and preserving the integrity of MS data.

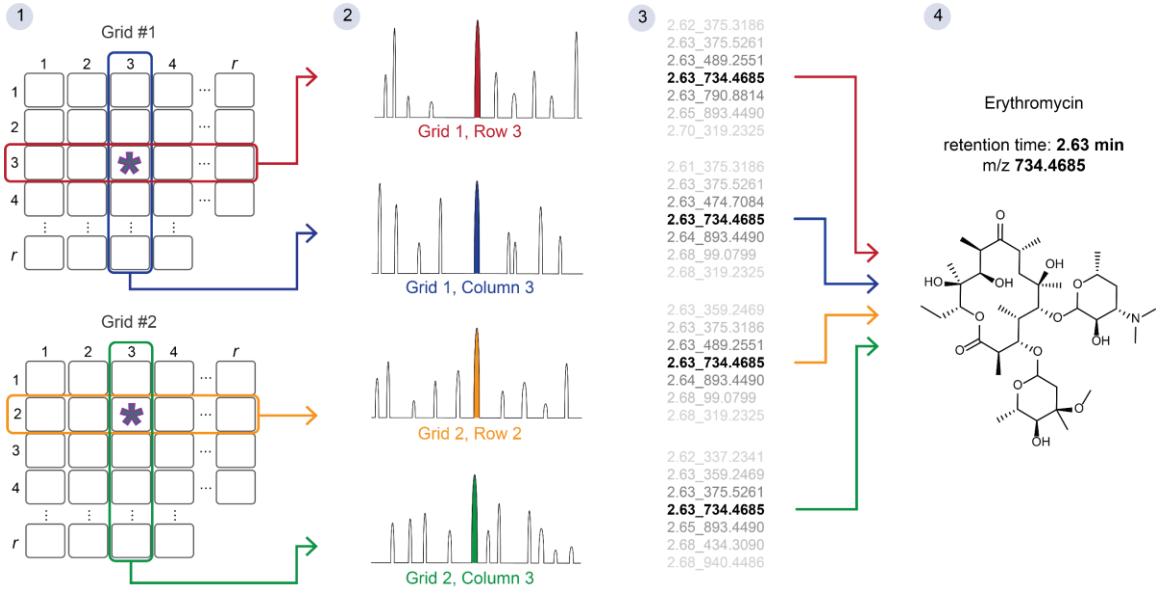


Figure 2.2 **Overview of the MultiplexMS Organization (MMSO) Strategy.** Symmetrical initial and rearranged grids of size r^2 are generated using the MultiplexMS app. Rows and columns of each grid are individually pooled and analyzed by UPLC−MS. Aligned MS features (retention time_*m/z* pair) in each dataset are traced back to the intersecting well of the analyzed row/column. If the sample at the intersection is the same in both grids, then the feature is assigned to that sample.

The pooled samples then undergo analysis via UPLC-MS. This is followed by the generation of a feature list for each mixture using a user-selected software package such as MZmine 3,[97] Waters Progenesis QI, etc. Subsequently, MMS employs computational deconvolution to decipher the initial and rearranged grids. This involves assigning features detected at the intersection of row/column combinations in both grids to the corresponding sample position. A feature must be detected accurately in matching samples across both grids to be added to that sample feature list. One notable advantage of this method is the built-in replicate comparison for quality control, as each

sample undergoes analysis four times (each row/column in both grids), ensuring robustness in results (**Figure 2.2**). The MMSO approach is specifically crafted to increase confidence in analyte assignments, particularly in scenarios where the same analyte appears multiple times within a sample set—a potential concern associated with orthogonal sample pooling.[84]

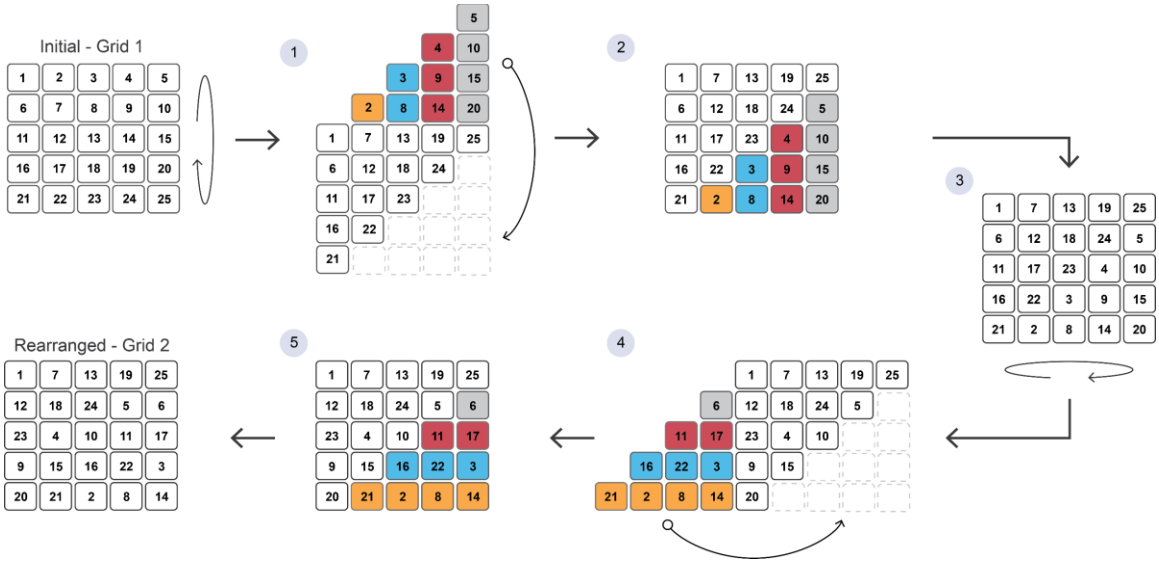


Figure 2.3 **The MultiplexMS Organization (MMSO) Protocol.**
(**1**) Starting with the initial grid (plate 1), the columns are rotated upwards incrementally, (**2**) The displayed samples are replaced into the grid in the new location, (**3**) and (**4**) The rows of the new grid are then rotated left incrementally, (**5**) The displayed samples are replaced to the new location to make the rearranged grid (plate 2).

2.2.2. *In Silico* Testing of the MultiplexMS Strategy

The MMSO sampling strategy and deconvolution protocol in MMS can pose a challenge when the same feature(s) appear across multiple samples within a grid. This can result in erroneous features being assigned to additional samples. Conversely, if the NP libraries are primarily composed of infrequent molecules, the deconvolution algorithm can accurately reconstruct MS feature lists for each sample. To rigorously evaluate the efficacy of the MMSO method, we conducted an *in silico* experiment using LC-MS data from 1,015 bacterial pre-fractionated extracts sourced from our in-house library. This dataset was chosen based on the taxonomic relatedness of the source organisms (Actinobacteria), and the likelihood of NP overlap between samples, providing an ideal test scenario to quantitatively assess the performance (precision and FPR) of the MMSO method for NP mixtures (**Appendix Figure A.2**).[27,98]

To estimate feature frequency in metabolic profile datasets, we performed an *in silico* subsampling of the full dataset and counted feature presence for all MS features. This experiment was repeated 50 times for three different grid sizes (r = 10, 20, and 30), containing 100, 400, and 900 prefractions, respectively (**Figure 2.4**). The number of occurrences, x, for each MS feature was averaged for all repetitions and plotted for each grid size (**Figure 2.5**). In all three cases, the average frequency of an MS feature being present just once exceeded 2,000 (r = 10, 2,654 ± 391; r = 20, 2,273 ± 149; r = 30, 2,040 ± 71) out of an average of 10,253 ± 657, 13,686 ± 251, and 15,453 ± 98 total MS features in each subsampling dataset, respectively. The frequency of MS features present twice in the dataset (x = 2) surpassed 1,000 for all grids (r = 10, 1,380 ± 219; r = 20, 1,258 ± 119; r = 30, 1,196 ± 30), decreasing for higher frequency counts (e.g., x = 5: r = 10, 544 ± 189; r = 20, 571 ± 97; r = 30, 512 ± 36).

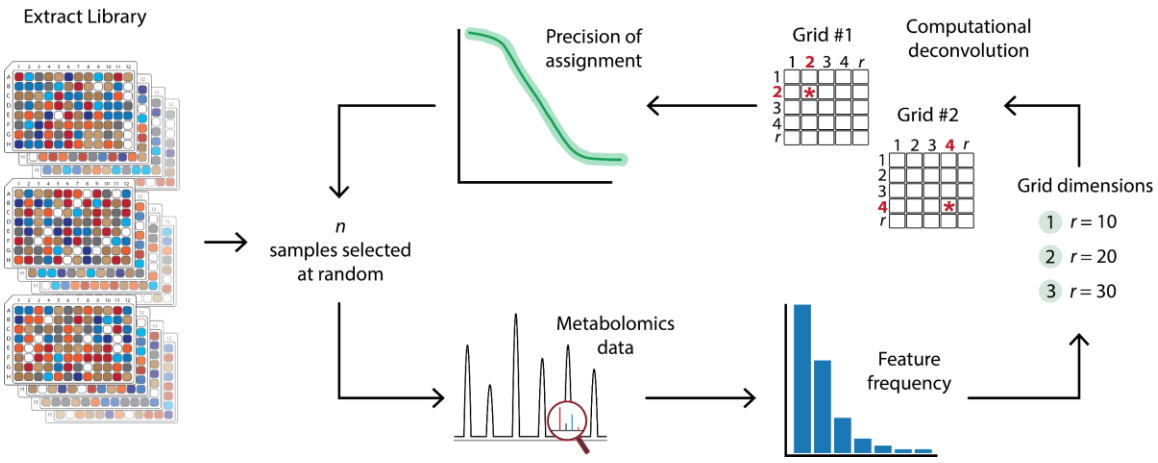


Figure 2.4 ***In Silico* Testing Scheme of the MultiplexMS Strategy.** The *in silico* testing scheme to test the validity of MultiplexMS. From a pool of 1,015 marine bacteria prefractionated library, subsets of size n = 100, 400, and 900 were randomly drawn 50 times with sample replacement. The size of the subsets represents grid sizes of r^2, where r = 10, 20, and 30. The number of occurrences, x, for each MS feature, was averaged for all repetitions for each subset size (blue bars in **Figure 2.5**, **Appendix Figure A.3**). Next, the subsets were arranged in MMSO-configured grids and computationally deconvoluted to determine assignment precision (green trendline in **Figures 2.5**, **Appendix Figure A.3**).

Additionally, we assessed the precision and FPR of the MMSO method for the three selected r^2 grid sizes (**Figure 2.5**). The MMS deconvolution algorithm was applied to simulate pooled rows and columns *in silico*, generating MS feature lists for reconstructed samples. Despite limitations such as excluding effects that could lead to

feature loss (i.e., ion suppression, sample preparation issues), this experiment provided valuable insights into the computational deconvolution algorithm performance under ideal conditions.

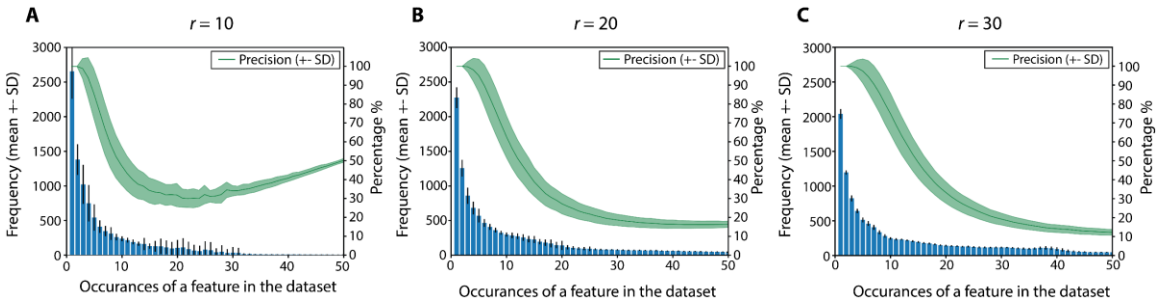


Figure 2.5 ***In Silico* Testing of the MultiplexMS Strategy.**
The results of the MS feature frequency count and precision calculation for each subset population using the dual grid sampling scheme with (**A**) r = 10, (**B**) r = 20, and (**C**) r = 30 grid dimensions.

The results showed that most MS features were present five times or fewer (e.g., r = 10, 62%), illustrating the sparse distribution of MS features in the prefraction library. Moreover, the MMSO sampling strategy improved assignment precision across all experimental grid sizes (e.g., x = 5: r = 10, 86 ± 13%; r = 20, 95 ± 9%; r =30, 97 ± 6%) compared to the single-grid approach (x = 5: r = 10, 31 ± 8%; r = 20, 25 ± 5%; r = 30, 23 ± 4%, **Appendix Figure A.3**). However, a balance must be maintained between throughput and FPR as grid dimensions r^2 increase. The FPR is inversely proportional to grid size r, with precision declining as feature frequency increases. Nonetheless, for drug discovery applications targeting "rare" metabolites sparsely distributed in a library, metabolites occurring with high frequency and weakly correlating with biological activity can be safely excluded from downstream analyses, mitigating concerns about FPRs for abundant features.

2.2.3. MultiplexMS Successfully Deconvolutes “One-Compound-One-Well” Libraries

Virtual experimentation (**Section 2.2.2**) highlighted the potential of the MMSO strategy to sample and deconvolute complex sample libraries successfully. However, the methodology does not consider real-world practical issues (e.g., ion suppression, pipetting errors) that can impact MS data quality. To demonstrate real-world scenarios beyond the *in silico* trials, increasingly intricate experiments were designed to gauge the

method's accuracy under practical conditions. As a proof-of-concept, we organized a set of 25 commercially available antibiotic and antifungal NP standards (n = 25) in MMSO grids (r = 5) following a “one-compound-one-well” setup (**Appendix Figure A.4**). This straightforward multiplexing scheme served as a litmus test to reveal any potential pitfalls like sample pooling effects during MS analyses, caused by instances such as chromatographic overlap and ion suppression.

The MMS app generated the initial and rearranged grids from the provided sample list of laboratory standards. Pooled samples were prepared by combining appropriate rows and columns within each grid, resulting in 20 pooled samples in total. The combined samples were diluted to an appropriate final concentration and subjected to UPLC-MS analysis. Preprocessing of the MS data involved peak picking and alignment of MS features using Progenesis QI, with a strict blank subtraction and minimum intensity threshold applied to the output file (see **Section 2.4.3**). Subsequently, MS feature lists derived from each pooled sample were computationally deconvoluted using the MMS “Deconvolute” function to reconstruct feature lists for each NP standard. Simultaneously, each of the 25 NP standards underwent individual MS analysis to establish a ground-truth (GT) dataset comprising accurate *m/z*-retention time pairs. A thorough examination of the assigned *m/z*-retention time pairs revealed that all 25 NPs were correctly assigned to the reconstructed samples, barring two instances of false-positive assignments (**Appendix Figure A.5**). In one case, $[M+H]^+$ *m/z* 335.10 was correctly assigned to penicillin G but incorrectly identified in the reconstructed sample of cloxacillin. Closer examination attributed this to low-intensity contamination of the commercial cloxacillin standard with penicillin G (**Appendix Figure A.6**). The second incorrect assignment involved the structural isomers tetracycline and doxycycline, where subpar peak shapes and overlapping elution times complicated the automated peak picking and alignment (**Appendix Figure A.7**). Overall, the precision of feature assignments was 96%, impaired by two instances of false-positive assignments. Encouraged by this promising pilot, the methodology was expanded to tackle the analysis of complex mixtures and delve into the detection limits for MS-based complex mixture analysis.

2.2.4. MS Feature Deconvolution Performance for LC-MS Analysis of Complex Mixtures

One of the primary concerns regarding multiplexing is the potential for analyte interference, which could result in the loss of information compared to individual sample analysis. Factors such as ion suppression, chromatographic peak overlap, and the limitations of automated peak detection software can complicate detecting MS features in complex mixtures. To understand how increasing sample complexity affects MS feature recovery, we conducted a study using a prefraction from our marine bacterial library. This prefraction, labeled as *i*, contained various NP compound classes, including micromonolactam (**2.1**), dracolactam A (**2.2**), and dracolactam C (**2.3**).[27,99,100] The prefractions were analyzed in triplicate, and retained MS features in all three replicates established a GT feature list.

Prefractionated extracts from other source organisms were sequentially added to the target prefraction, starting from *i* + *1*, *i* + 2,..., up to *i* + 49, and then incrementally for *i* + 59, 69, 79, and 99. Each set of pooled prefractions represented potential grid dimensions from *r* + 1 to 100 (**Figure 2.6**). The concentrations of individual prefractions in each pooled sample of size *r* were kept constant across the set to eliminate dilution effects. Each pooled sample was analyzed by UPLC-MS, preprocessed in Progenesis QI to generate an MS feature list, and compared to the GT feature list to determine recovery rates for the GT MS features from the target prefraction.

Feature recovery of the target prefraction was assessed in two ways. First, the three benchmark molecules (**2.1** – **2.3**) were examined to determine information recovery in the presence of increasing numbers of additional prefractions (**Figure 2.6.B**). Extracted ion chromatograms (EICs) for each molecule demonstrated consistent peak shape and intensity up to *r* = 60. Beyond *r* = 70, although peak shapes remained consistent for compounds **2.2** and **2.3**, peak intensities decreased, reaching a minimum of 64% for **2.2** and 43% for **2.3** of the GT signals when *r* = 100 (**Figure 2.6.B**). Similarly, for benchmark molecule **2.1**, peak shape remained consistent through *r* = 100, but peak intensity decreased at *r* = 70, with the most significant reduction in intensity observed at 50% of the GT signal for each molecule when *r* = 100 (**Figure 2.6.B**).

Secondly, the recovery of the complete GT feature list from the target extract was analyzed for each test mixture. The percentage of recovered features from the GT

feature list plotted against mixture complexity ($r = 0 \rightarrow 100$) showed that pooling up to 10 extracts returned over 97% of GT features. However, increasing the mixture complexity to 20 samples reduced recovery to 90%, while pooling up to 30 samples decreased the overall recovery rate to 74% (**Figure 2.6.C**). Analyzing signal intensities for unrecovered features indicated that most were low-intensity analytes near the intensity threshold (**Figure 2.6.D**). Since most NPs possess multiple MS features under standard LC-MS acquisition conditions, the loss of a few low-intensity features is unlikely to significantly impact the chemical characterization of complex mixtures, especially for drug discovery applications.[29]

An essential insight from this analysis is that grid size selection should not solely prioritize reducing acquisition time but should instead consider optimizing sample complexity and MS feature recovery. The complexity, chemical similarity, and chromatographic properties of the mixtures under analysis heavily influence these factors. Users of the MMS platform are recommended to perform the outlined benchmarking method with their sample libraries and refer to the plot in **Figure 2.6.C** to select an appropriate grid size for each sample set.

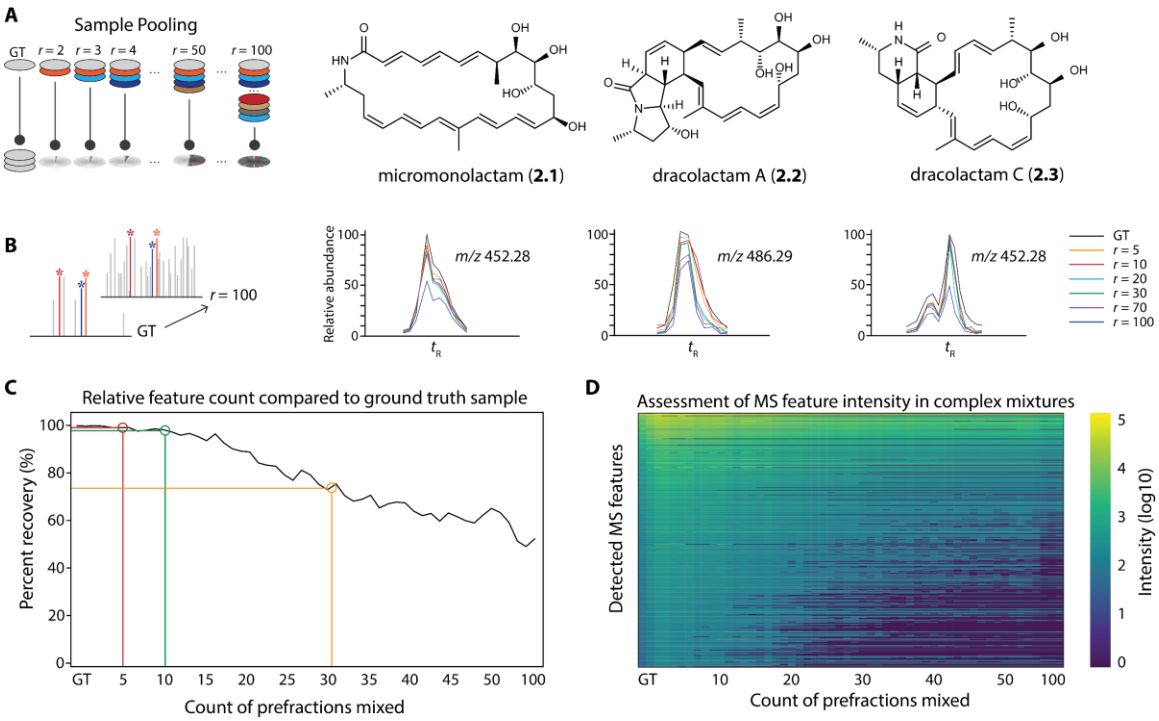


Figure 2.6 **MS Feature Deconvolution Performance for UPLC-MS Analysis of Complex Mixtures.**
(**A**) Pooling scheme showing the successive addition of complex mixtures to a GT sample containing known molecules **2.1**, **2.2**, and **2.3**. (**B**) Information recovery of benchmark molecules. Each molecule shows consistent peak shape and intensity up to $r = 60$ in all cases, while signal intensity reduces above $r = 70$. (**C**) Relative feature recovery of the full MS feature list from the GT feature list was assessed for each test mixture and plotted as percent recovery. Red: $r = 5$, Green: $r = 10$, Yellow: $r = 31$. (**D**) Assessment of feature recovery as a function of mixture complexity. GT MS features are on the y-axis in order of feature retention success throughout the dataset.

2.2.5. Assessment of MultiplexMS Performance with Complex Mixtures

The assessment of MS feature recovery concerning sample mixture complexity revealed a return rate exceeding 97% for mixtures comprising ten extracts (**see Section 2.2.4**), indicating the proficiency of the MMS method in medium-sized grid setups. However, this evaluation overlooked alterations in the MS feature list composition due to factors like ion suppression and peak overlap. To gauge the method's performance in a real-world scenario, MMSO grids housing 90 bacterial prefractions and 10 NP laboratory standards were generated so that each row/column in the initial grid contained one NP standard (**Appendix Figure A.8**). This arrangement provides an internal reference during the deconvolution step. After pooling and analyzing rows and columns via UPLC-

MS (totaling 40 samples) per the MMSO protocol, each prefraction was independently analyzed in quadruplicate to establish a strict GT MS feature list.

Preprocessing of the multiplexed MS data in Progenesis QI, involving blank subtraction and enforcing a minimum intensity threshold, pooled samples underwent computational deconvolution to regenerate MS feature lists for each original prefraction. NP standard-associated MS features were accurately linked to their respective samples (**Appendix Figure A.9**) without any false-positive assignments. Next, MS feature lists from the GT prefraction replicate analyses were compared to the reconstructed samples from the multiplexed plates. Feature recovery was quantified as the percentage of GT MS features found in the reconstructed feature list for each prefraction (**Appendix Figure A.10**). Recoveries spanned from a peak of 97% to a low of 57%, with 72% of samples boasting recovery rates of 80% or higher. Notably, most samples with diminished feature recoveries (<70%) comprised nonpolar “wash” fractions containing relatively sparse MS features in the GT dataset (**Appendix Figure A.11**). The slight feature loss in these cases resulted in a considerable reduction in percentage feature recovery due to the small denominator value. As observed in the previous pooling experiment, absent features were predominantly low-intensity *m/z* ions, high-frequency background ions, isotopologues, or multiply charged ions identified by Progenesis QI. Despite these omissions, crucial for precise atomic composition determination, they did not impact the informational integrity of the reconstituted MS feature lists for associating molecule presence/absence with specific biological traits.

This experiment supported the outcome aligned with the *in silico* analysis (**Figure 2.2**), affirming excellent FPRs across all samples but high rates for commonly occurring features. Using a 10 x 10 grid curtailed data acquisition time tenfold compared to quadruplicate individual analyses, with minimal data loss compared to GT MS feature lists. The considerable overlap in feature recovery between *in silico* and experimental analyses lends credence to this approach for bioactive compound discovery using automated data integration methods.

2.2.6. Application of MultiplexMS to Ultra-High-Throughput Library Analysis

Prioritizing bioactive compounds is crucial in NP drug discovery, mainly when only a small subset of analytes in NP extract libraries display bioactivity.[86] Metabolomics data, alongside bioassay data collected from a diverse set of NP extracts, can effectively spotlight active constituents, streamlining the discovery process in NP pipelines. NP Analyst, an open platform for Compound Activity Mapping (CAM), serves as a tool integrating metabolomic and bioactivity data to facilitate compound prioritization.[27] In the original study from our laboratory by Lee *et al*., metabolomics data from 925 in-house marine bacteria prefractionated extracts (155 source organisms), measured in technical triplicates (amounting to 2,775 samples), coupled with an Antibiotic Mode of Action Profile (BioMAP) screening panel, were leveraged to pinpoint priority molecules for isolation, ultimately leading to the discovery of new bioactive compounds.[27,101]

The acquisition of MS data for the 925-member prefraction library demanded over 15 days of continuous acquisition time, effectively tying up a high-resolution (HR) mass spectrometer for several weeks, significantly longer than the time required for acquiring the corresponding biological screening results. This original NP Analyst experiment provided an ideal testing ground to benchmark the MMSO strategy against a GT dataset. Since metabolomics data had already been acquired for each sample in triplicate in the original NP Analyst study, the principle of the MMSO strategy was initially tested by pooling feature lists *in silico* using an *r*-value of 31 (31 x 31). This grid size of *r* = 31 is the minimum size required to accommodate all 925 samples in a single analysis, offering the highest possible throughput. This approach reduced the total number of samples needed for MS analysis from *n* = 2,775 (925 × 3) to 124. It reduced the theoretical analysis time from 15 consecutive days to just 15 hours, marking a 24-fold reduction in acquisition time.

NP Analyst prioritizes MS features based on the activity profile strength (Activity Score) and consistency (Cluster Score) for the samples where the MS feature is found.[63] Errors in reconstitution from MMS could affect both Activity and Cluster scores, potentially deprioritizing essential bioactive molecules. To assess the value of MMS data for bioactive compound discovery, a new NP Analyst experiment was conducted using the reconstituted MS feature lists from the *in silico* MMS experiment and the BioMAP

activity data from the original NP Analyst study. Subsequently, the Activity and Cluster Scores for each MS feature were compared between the original and *in silico* MMS experiments. Changes in Activity and Cluster scores for active features from the original experiment provided an objective measure of the influence of MMS analysis on the bioactive compound discovery, the primary application for which this platform was designed.

The original metabolomics dataset contained 9834 detected MS features, of which 845 were defined as 'active' using Activity and Cluster score filters of 2.0 and 0.3, respectively. Following computational deconvolution of the *in silico* pooled dataset, the Activity and Cluster scores were recalculated on all active and inactive MS features. An MS feature presence filter was applied, eliminating features in >20 reconstructed prefractions (**Figure 2.7** and **Appendix Figure A.12**). Encouragingly, all 845 bioactive MS features from the original experiment were detected in the filtered virtual experiment. Of these, 798 showed no change in Activity and Cluster scores, indicating correct assignment to their original positions in each grid. Among the 47 features with changes in either Activity or Cluster Score, changes ranged from -4.1 to +0.23 (Activity Score) and -0.46 to +0.042 (Cluster Score) (**Figure 2.7.B, C**). In total, only 26 features had changes in scores large enough to change their assignment from 'active' to 'inactive.' Notably, there were no 'active' misassignments in the *in silico* NP Analyst dataset, demonstrating high assignment precision and zero false positives.

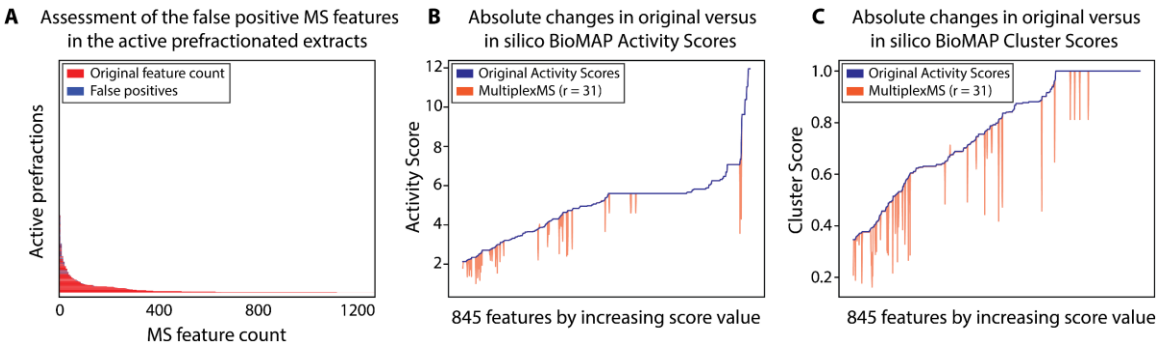


Figure 2.7 ***In Silico* MultiplexMS Comparison of Activity and Cluster Scores to a GT Experiment Set.**
(**A**) Assessment of the FPR in the active prefractions following *in silico* MultiplexMS. First, MS features present in the dataset ≥ 20 samples were omitted. Next, an Activity and Cluster Score filter of 2.0 and 0.3 were applied. (**B**) Absolute changes in the original Activity Scores of the active MS features compared to the *in silico* MultiplexMS experiment. (**C**) Absolute changes in the original Cluster Scores of the active MS features versus the *in silico* MultiplexMS deconvoluted scores.

The entire NP Analyst analysis was repeated experimentally as a final validation by creating pooled samples from the original stock solutions and reacquiring the raw MS data. The prefraction sample list was provided to the MMS app to generate the initial and rearranged grids for $r = 31$ and create the pooled lists for the rows and columns of each grid. Pooled samples were analyzed via MS and preprocessed in Progenesis QI for peak picking and alignment, and the pooled feature lists were computationally deconvoluted. The comparison between the original and deconvoluted datasets revealed the extent of feature inflation when pooling up to 31 prefractions, providing an ideal test case for prioritizing the rarer analytes in a dataset.

The prefraction MS feature lists and the original BioMAP dataset were integrated using the NP Analyst platform, generating a prefraction-feature activity network for bioactive compound identification and prioritization (**Figure 2.8** and **Appendix Figures A.13 – A.15**). Encouragingly, the new analysis recapitulated the creation of distinct communities for the known bioactive molecules from the original study (**Figure 2.8**). These included communities for micromonolactam (**2.1**) and dracolactam A (**2.2**); the collismycin analogs collismycin A (**2.4**), collismycin B (**2.5**), and SF2738D (**2.6**); and amychelin C (**2.7**). The characterization data for each molecule were compared to the previous report, including *m/z*, retention time, Activity Score, and Cluster Score for representative ions from each molecule (**Figures 2.8.B**). All previously identified molecules were present in the MMS version of the experiment, displaying remarkably similar chromatographic characteristics and instrument response. Moreover, the Activity and Cluster scores remained identical in nearly all instances.

These results demonstrate that MMS can successfully decrease acquisition time and increase sample throughput for bioactive compound discovery. In a complex real-world example containing over 900 samples, the system performed equivalently to the gold-standard approach with individual replicate analyses but required less than 5% of the MS acquisition time. This outcome suggests that, at the compound level, the information content was equivalent, paving the way for application in other large NP-based screening projects where data acquisition rates currently limit the use of next-generation data integration strategies like NP Analyst.

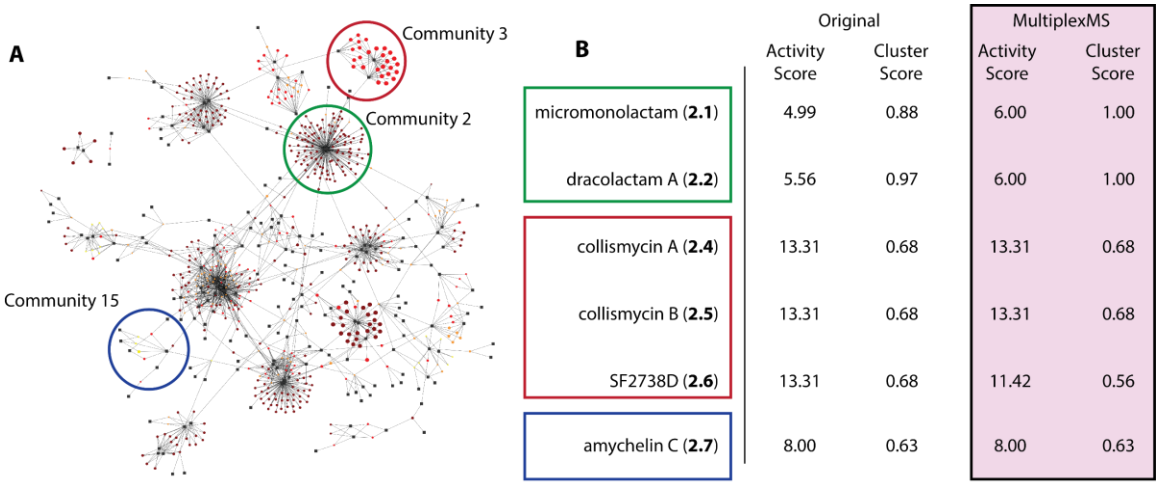


Figure 2.8 Ultra-High-Throughput MultiplexMS Application with Complex Mixtures.
(**A**) NP Analyst output from multiplexing 925 prefractionated extracts in 31 x 31 sampling grids (124 pooled samples). Distinct communities of MS features were generated. Circled communities highlight dereplicated molecules in the dataset. (**B**) Activity and Cluster Score comparison between molecules identified in the original analyses versus MultiplexMS.

2.3. Conclusion

Advancements in instrument sensitivity and preprocessing software are key in guiding the identification of bioactive compounds in high-throughput screening applications. However, as laboratory automation progresses, the scale of NP libraries is increasing substantially.[20] Consequently, the time required to analyze these samples by LC-MS has surged, sometimes spanning decades of instrument time in extreme cases. Complicating matters further, many commercial and third-party MS processing software packages struggle to handle peak picking and alignment of thousands of samples in a single analysis, rendering traditional methods impractical.

Throughput is a recognized bottleneck in metabolomics, prompting various strategies, such as pooled quality control (pooled QC) samples to define complete dataset peak lists, to address this challenge.[102] This acquisition strategy can analyze thousands of samples by utilizing the total peak list from a pooled QC sample to select batches of research samples.[89] However, despite its effectiveness in eliminating the need for technical replicates for all samples, the number of MS runs still increases linearly with the sample count. This contrasts MultiplexMS, which requires significantly fewer MS runs for a given sample set.

In metabolomics, pooled QC works well for samples with similar compositions but varying analyte concentrations. However, pooled QC is less effective in bioactive compound discovery, where compositions vary widely due to signal suppression and overlap. MMS focuses on detecting feature presence/absence rather than quantification, allowing for increased throughput. One drawback of this approach is that FPRs can be elevated for commonly encountered features. For bioactive compound discovery, this is of limited concern. Sampling strategies in natural product research often prioritize maximizing species[103] and chemical diversity[104] within the sample set to uncover "rare" metabolites with unique structural and biological properties. Hence, commonly encountered features can be disregarded due to their weak correlation with biological phenotypes, as demonstrated by the NP Analyst experiment (**Section 2.2.6**). To mitigate false positives, users can design chromatographic methods that minimize overlap and incorporate appropriate blank subtraction processes into workflows. This eliminates prevalent background MS features from multiplexed rows and columns, ensuring peak lists accurately represent mixture composition. Additionally, employing HR-MS systems with resolving powers exceeding 20,000 and considering additional separation axes such as ion mobility spectrometry (IMS) can reduce the occurrence of incorrect feature alignment of features with similar physiochemical properties.

In summary, MMS offers an HT pipeline for swiftly acquiring qualitative insights into the chemical complexity of large NP libraries, facilitating the exploration of extensive collections. Supported by an open-source software package with a user-friendly GUI, MMS is accessible to non-programmers and vendor-independent, seamlessly integrating into existing MS workflows.

2.4. Materials and Methods

2.4.1. General Experimental Information

All solvents used in the mass spectrometry acquisition were of optima LCMS grade. Acetonitrile (ACN), methanol (MeOH), and formic acid (HCOOH) were purchased from Thermo Fisher Scientific. Deionized water was obtained using an 18 MΩ·cm Milli-Q system (EMD Millipore Corporation). Clindamycin, deferoxamine, erythromycin, novobiocin, nystatin, roxithromycin, and tetracycline were purchased from Sigma Aldrich. Azithromycin was purchased from TCI. Lincomycin was purchased from RPI. Mupirocin

was purchased from Applichem. Staurosporine was purchased from LC Laboratories. HR-MS acquisition and metabolomics experiments were performed on a Waters SYNAPT G2-Si qTOF mass spectrometer. NMR spectra were measured on an AVANCE II 600 MHz spectrometer with a 5 mm QCI cryoprobe (Bruker).

2.4.2. UPLC-MS Acquisition Conditions

All measurements were performed on an ACQUITY UPLC *i*-Class system (Waters) with an ACQUITY HSS T3 column 1.8 µm (2.1 x 100 mm, Waters). Separation was achieved using a linear elution gradient (mobile phase A: H_2O + 0.01% formic acid; mobile phase B: ACN + 0.01% formic acid) as follows: 0 – 0.3 min, 5% B; 0.3 – 4.7 min, 5% – 90% B; 4.7 – 5.5 min, 90% – 98% B; 5.5 – 5.8 min, 98% B; 5.8 – 7.5 min, 5% A. The flow rate, column temperature, and injection volume were set to 0.5 mL/min, 40 °C, and 5 µL, respectively. MS data were acquired on a SYNAPT G2-Si Quadrupole-Time-of-flight (qTOF) mass spectrometer (Waters) equipped with an electrospray ionization (ESI) source. MassLynx v.4.1 SCN941 was used as the instrument acquisition software. Mass measurements were recorded using ESI+ data-independent acquisition (DIA) experiments. Detection was performed in the *m/z* range of 50 − 1500 with a scan rate of 5.0 Hz, a capillary voltage of 3.0 kV, and a desolvation temperature of 300 °C. Leucine enkephalin (Waters SKU: 186006013) was employed as the lock spray solution at 200 pg/µL concentration at 0.10 Hz.

2.4.3. Mass Spectrometry Data Processing

All samples were processed using the Progenesis QI software suite (v2.2.5826.42898, Nonlinear Dynamics, Waters). Non-lock mass corrected data was uploaded into Progenesis QI software for spectra alignment, lock mass calibration (applied to the standard *m/z* 556.2771), and peak picking using the default settings. The generated feature table, including the *m/z*, predicted neutral mass, retention time, and intensity values, were exported, and provided to the MMS software for computational deconvolution. A strict blank subtraction was implemented whereby MS features in the solvent blank samples were omitted from the resulting MS feature table. Then, a minimum intensity threshold was applied to the MS features in the table, eliminating features from a sample that fell below the threshold. The minimum intensity threshold

value was determined according to the *Determination of minimum intensity threshold* protocol (**see Section 2.4.4**).

2.4.4. Determination of Minimum Intensity Threshold

The metabolomics dataset of 1015 prefractions and methanol/water blank measurements underwent preprocessing in Waters Progenesis QI using default settings. We calculated the feature count per sample across a range of intensity cut-off values to identify a cut-off threshold below which a feature was considered absent. We reasoned that feature counts would significantly increase below a certain intensity cut-off threshold, indicating the presence of noise features in the dataset. We applied a cut-off threshold of 70, determined from the inflection point in the feature count–threshold intensity graph (**Appendix Figure A.16**). Subsequently, features below this cut-off value were set to 0, and any features present in at least 50% of the blank samples were removed from the dataset.

2.4.5. Sample Preparation

Schulze *et al*.[98] outlined the process for bacterial strain isolation, culture, extraction, library preparation, and crude extract fractionation. In this study, aliquots from designated wells in DMSO stock plates containing prefraction samples were transferred into Corning V bottom 96-well plates. These aliquots were diluted 1:100 in DMSO, and the replicate parent plates were stored at -70 °C for future use. The arrangement of samples in the multiplexed destination plate was determined using the “random” function within MMS. Based on the sample library size *n*, MMS will produce r^2 symmetrical grids to accommodate all samples in the library where

$$grid\ count = \left\lceil \frac{n}{r^2} \right\rceil \qquad \textbf{(Equation 2.1)}$$

In the final validation experiment, a library size of n = 925 with grid dimensions of r = 31 generates a single initial and rearranged grid. Due to the MMSO methodology, the total number of pooled samples equals 4r x grid count.

Sample mixing was carried out using a TECAN Evo 150 liquid handler equipped with a LiHa robotic arm for automated multiplexing. In the NP Analyst recreation

multiplexing experiment, 20 µL was removed from the parent plate well, and 5 µL was dispensed into assigned well coordinates in the initial and rearranged destination plates. This process was repeated for all experimental wells until the multiplexing grids were complete. Once finished, the destination multiplexed plate was dried *in vacuo* to remove the DMSO. The dry residue was then resuspended in 5 µL DMSO (1:100) and diluted with 1.5 mL 50% (v/v) methanol/water (1:30,000) for mass spectrometric analysis.

2.4.6. MultiplexMS Application Development and Pipeline

The MMS application is available for both Windows and MacOS operating systems. It is supported by extensive online documentation at liningtonlab.github.io/MultiplexMS_documentation. MMS requires only preprocessed MS1 level MS data in the deconvolution pipeline. This data consists of a multiplexed dataset's individual MS features (*m/z* vs retention time pairs). The flexibility of MMS allows it to work with a wide range of data acquisition modes, making it compatible with various experimental setups. The MMS application simplifies the user process. First, users upload a sample list to the platform. Then, MMS automatically arranges these samples in user-defined, symmetrical grids or *r* x *r*. Next, the user prepares the pooled samples based on row and column combinations for MS analysis. The feature table is returned to the MMS application after preprocessing the MS data using their preferred software package. Here, the rows/column mixtures are computationally deconvoluted back into MS feature lists for each original sample in the set. The app also includes data validation and quality control checks, guiding users on necessary corrections before submitting their jobs. It is divided into two main sections: preparation and deconvolution.

i) Preparation Page: MMS offers sample preparation utilities that help organize samples into symmetrical grids of length *r*. This feature allows users to customize the pool size and define how samples are organized within the pools (whether in a random or fixed order). To execute the preparation step, users must select the file containing the sample names and specify the output directory for the resulting plate maps and preparation tables. The plate organization between the initial and rearranged multiplexed plates adheres to the MMSO protocol (**Figure 2.3**).

ii) *Deconvolution Page*: After conducting MS analyses and preprocessing the data, users choose plate maps from the preparation stage and MS feature lists for pooled samples. The MMS tool then uses these plate maps as coordinates to reconstruct samples from the pooled sample feature lists. The reconstructed feature lists are saved to a predefined output folder. Optionally, users can remove features that appear in more than a specified number of samples using the 'Critical threshold' parameter in the Cleaning tab after deconvolution. By default, the tool excludes MS features present in the pooled sample MS data but not associated with any individual samples. However, users can deselect this exclusion option if they need a complete list of all MS features in the pooled sample feature lists (e.g., for examining the fate of specific MS features).

2.4.7. *In Silico* Subsampling and Feature Frequency Estimation

The *in silico* subsampling approach was used to evaluate the mass spectrometric feature frequency and precision of the MMS method on a metabolomics dataset of 1015 prefractions. After processing the metabolic profiles in Waters Progenesis QI with default settings, we obtained a flat feature list with features represented as *m/z*–retention time pairs. Additionally, deconvoluted analytes were represented by their molecular mass (M) and corresponding retention time. To enhance data quality, a minimum intensity threshold (> 70) was applied to remove noisy features from the dataset (see **Section 2.4.4**). Subsequently, the dataset was binarized (1 for feature present, 0 for feature absent). We subsampled the overall dataset to estimate feature frequency based on a selected r^2 grid (e.g., $r = 10$). For each subsample, we tallied the presence of MS features across the entire set. This subsampling process was repeated 50 times, and the resulting feature frequency values were reported as averages along with standard deviations.

To simulate the MMS method, we followed the following steps: i) randomly draw a subsample of size n and arrange the samples in a square grid with dimensions r x r (where $r = 10$, 20, or 30) to mimic the mixing process.; ii) obtain the presence vector for each feature in the square grid (length r x r) to generate the initial presence vector (GT); iii) apply the MMS strategy to create a rearranged presence vector; iv) Construct a new observed vector representing the multiplexed samples using the initial and rearranged vectors; v) determine the deconvoluted presence vector based on overlapping feature

presence using the observed presence vectors; vi) calculate precision by comparing the GT vector with the deconvoluted vector. The precision calculations were measured as

$$precision = \frac{TP}{TP+FP} \qquad \textbf{(Equation 2.2)}$$

where TP = true positive and FP = false positive. Precision was used to measure the repeatability of each iteration versus the instances of false positives. The FPR is calculated as 1 - precision. The *in silico* MMS scheme used a dual-grid and single-grid orthogonal strategy to determine MS feature assignment accuracy between methods.

2.4.8. Isolation, Fermentation, Extraction, and Fractionation

Frozen stocks were inoculated on solid media (10.0 g of glucose, 5.0 g of NZ-amine, 1.0 g of $CaCO_3$, 20.0 g of starch, 5.0 g of yeast extract, 20.0 g of agarose, and 1.0 L of water) and incubated at room temperature until discrete colonies became visible. Individual colonies were inoculated into seed cultures in 40.0 mL culture tubes with 7 mL of SYP culture medium (31.2 g of Instant Ocean, 10.0 g of soluble starch, 4.0 g of peptone, 2.0 g of yeast extract per 1 L of water). After four days, 3 mL of the liquid culture was used to inoculate 60.0 mL of SYP in 250 mL Erlenmeyer flasks. The culture was incubated for five days before 30 mL was transferred to large-scale, 1.0 L media in 2.8 L Fernbach flasks along with 20.0 g Amberlite XAD-16 absorbent resin and a large stainless-steel spring. Large-scale cultures were fermented for seven days, filtered using glass microfiber filters, and washed with water (500 mL). The combined cells and resin were extracted with 1:1 DCM/MeOH (500 mL), filtered, and the filtrate concentrated to dryness *in vacuo*. Then, the dried extract was fractionated by solid phase extraction (SPE) using a MeOH/H_2O stepwise LC gradient to afford six prefractions (A – F).[98]

2.4.9. Isolation of Active Compounds from Prefraction RLUS-2152D

Compounds **2.1**, **2.2**, and **2.3** were initially isolated from RLUS-2152D after refermenting the producing organism. Due to supply limitations of the original prefraction, we included the new prefraction RLUS-2152E in this experiment. The newly fermented prefractions contributed additional molecules, increasing features within the NP Analyst network (Community 2). Interestingly, the original molecules persisted and displayed similar activity and cluster scores (**Figure 2.8**).

Community 3 showed the presence of **2.4**, **2.5**, and **2.6**. Compounds **2.4** and **2.5** exhibited identical activity and cluster scores, while **2.6** had slightly lower scores. The feature $[M+H]^+$ *m/z* 753.3066 at 2.30 minutes – corresponding to the parent mass of **2.7 –** was identified in RLUS-2105A, B, and C using Progenesis QI. This identification differed from the original assignment limited to B and C. Upon re-examining the deconvoluted prefraction, we confirmed the presence of the parent mass of **2.7**, albeit in very low abundance (still about the minimum intensity threshold applied to the dataset). Interestingly, prefraction RLUS-2105A did not induce activity across the panel of bacterial strains in the BioMAP screen. Consequently, the feature *m/z* 753.3066 received a low activity and cluster score. In contrast, the Fe-complex adduct of **2.7** maintained identical activity and cluster scores to the original scores (**Figure 2.8**).

Chapter 3.

MultiplexMS-Q: A Quantitative Mass Spectrometry Method for Relative Feature Abundance Determination in Multiplexed Mixtures

Manuscript in preparation

Authors: Michael J. J. Recchia, Tim U. H. Baumeister, Roger G. Linington

Author Contributions: M.J.J.R. developed and integrated the MultiplexMS-Q methodology into the existing application. M.J.J.R. conducted statistical analyses and qualitative assessments of the method's results. T.U.H.B. provided consultation on quantitative methodologies and assisted with custom scripts. Additionally, M.J.J.R., T.U.H.B., and R.G.L. collaborated on the manuscript's authorship and will review the final version before submission.

3.1. Introduction

Advancements in untargeted metabolomics analysis of natural product (NP) mixtures continue to evolve, facilitated by improvements in analytical instrument technology and computational processing.[105,106] These breakthroughs have significantly expanded our understanding of biological systems, enabling the exploration of intricate molecular landscapes. For example, enhancements in mass spectrometry (MS) resolving power enable the differentiation of two adjacent ions of equal intensity, allowing for the unequivocal determination of molecular formulae with ppb mass accuracy.[106–108] Liquid chromatography (LC)-MS has been a cornerstone in this field due to its exceptional sensitivity, enabling the detection of a vast array of metabolites in biological samples. However, the time-consuming nature of adequately separating molecules in complex mixtures remains a significant challenge in the discovery process. This bottleneck can hinder the identification of potential bioactive compounds, especially when analyzing large NP extract libraries.[18,36]

MultiplexMS (MMS) was developed as a solution to the throughput bottleneck, capitalizing on strategic sample multiplexing to reduce the time it takes to analyze large

sample libraries without compromising data quality.[109] This approach pools multiple samples in a single MS experiment and then computationally deconvolutes the preprocessed MS data to construct individual feature lists for each sample. The MMS algorithm produces a binarized presence/absence output (present = 1, absent = 0) that enables rapid analysis through computational deconvolution, offering scalability and efficiency for large-scale MS analysis of extract libraries in a single experiment.

While MMS offers significant advantages for high-throughput (HT)-MS analysis, the binarized methodology presents quantitative or targeted metabolomics limitations. These limitations arise from the need for relative concentration of metabolites across a broad sample set, which is essential for a comprehensive dataset analysis. For example, Lee *et al*. conducted a global meta-analysis of several heterologous datasets to assess metabolic signatures (diversity of signatures and metabolite abundance) corresponding to various phenotypes involved in several metabolic pathways.[110] The findings showed which metabolic pathways affected the abundance of metabolites for a given phenotype. Although nuclear magnetic resonance (NMR)-based metabolomics is also used, LC-MS allows various chromatographic columns to separate analytes in mixtures, reducing sample complexity and increasing sensitivity to detect many molecules simultaneously.[111,112] Metabolomics data is often presented as relative quantities of molecules across samples without using quantitative NMR or calibration curves for specific molecules.[113] The primary reason for this discrepancy is the varying ionization efficiencies of different metabolites within complex mixtures, which makes direct correlation with absolute concentrations challenging.[114,115] Furthermore, the impracticality of generating standard curves for potentially thousands of molecules exacerbates this limitation, posing a sizable obstacle in metabolomics research.[36] Determining the relative quantification values of a metabolite between samples offers advantages. Statistical comparisons of spectral features and intensities allow accurate identification of similar compounds and compound classes across sample sets, providing a comprehensive view of the metabolome and sample complexities.[116,117]

MultiplexMS-Q (MMS-Q) was created to address these challenges by generating relative quantitation data of MS features from pooled complex mixtures following MMS computational deconvolution. The computational framework MMS-Q uses matrix algebra and a least squares algorithm to calculate relative quantitative information about each retained MS feature based on the abundance value detected in a pooled sample. By

leveraging the MMS binary output, MMS-Q calculates abundance data based on the presence of detected features within an initial and rearranged grid pair. This methodology enables the determination of relative quantitation data for each detected MS feature in the dataset, which can benefit compound activity mapping (CAM) techniques that prioritize metabolites based on their relative abundance and potential functional relevance, helping to identify lead compounds or NPs with bioactive properties.[27,33,64,118]

This chapter describes the MMS-Q methodology, including integrating the quantitation tool into the existing MMS application, rigorous algorithm testing to optimize prediction accuracy, and potential applications in NP discovery.

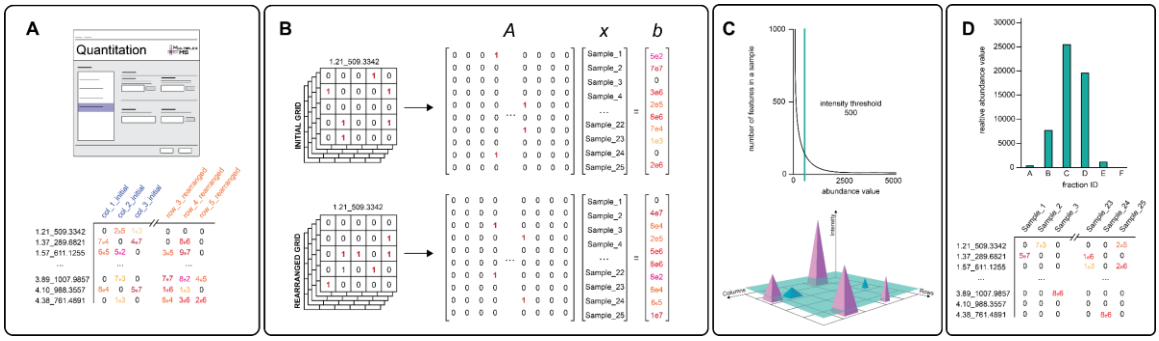


Figure 3.1 **Overview of the MultiplexMS-Q Workflow.**
(**A**) The MMS-Q quantitation app receives pooled MS data containing abundance values and binary data for each MS feature in the pooled sample. (**B**) The binarized data is mapped back to plate coordinates for each MS feature, and intensity information for each pooled sample is appended to the grids. These binary grids and intensity data are then organized into a matrix equation ($Ax = b$) to calculate the relative intensity value for the MS feature at its position in the grid. Each feature in the rows and columns contributes to the pool's abundance value and, therefore, carries a weight used to estimate the abundance value of the MS feature in the grid. (**C**) Researchers can determine the minimum intensity threshold by analyzing the abundance values in the table. Mapping the abundance values against the frequency of values determines an inflection point (minimum intensity threshold), which is then applied to the field of values. Any MS feature falling below this intensity value is set to 0 in the respective sample. (**D**) The relative abundance of a molecule can be weighed in each sample it was detected in, providing a comprehensive analysis of the metabolomics dataset.

3.2. Results

3.2.1. Implementation of the Relative Quantitation Component of MultiplexMS

The MMS-Q method for predicting relative quantitation values builds upon the multiplexing approach used in MMS.[109] MMS employs a dual-grid orthogonal pooling strategy to organize samples from a library of *n* samples in *r* x *r* grids (initial and rearranged). These samples are then multiplexed into pools of size *r*, with each sample contributing 1/*r* of the sample pool. The key innovation lies in the MultiplexMS Organization (MMSO) protocol, which strategically places samples between the two grids, ensuring that no two samples from the initial grid occupy the same row or column in the rearranged grid (**Figure 2.3**). Following ultra-performance liquid chromatography (UPLC)-MS, the pooled samples are computationally deconvoluted by assigning detected MS features to the sample occupying the intersection of the row/column combinations, generating feature lists for each sample in each grid. The sample feature lists are compared between the grids, and only the features present in both are retained. However, the binary nature of the MMS deconvolution process, while enabling rapid processing, limits relative quantitation analysis.

Preserving the abundance of data on MS features serves a dual purpose: enhancing relative quantitation analysis between samples and mitigating false positive assignments. During the computational deconvolution process in MMS, false positives become more common as the frequency of an MS feature within the grid increases. While detecting an MS feature in a single row and column pair is straightforward, challenges arise when multiple instances of the same feature exist (**Appendix Figure A.1**), creating a higher risk of false assignments. Although applying a minimum intensity threshold before MMS processing can reduce some instances, there may still be false detections that exceed this limit. The binary nature of the computational deconvolution method makes it difficult to confirm the presence of an MS feature accurately.

Relative abundance data can address this issue by comparing relative intensities across instances of an MS feature in multiple samples. This comparison enables the establishment of an additional threshold for determining the final peak list. The computational deconvolution algorithm also considers results from initial and rearranged

grid pairs. It requires a detected MS feature to be assigned to the same samples in each grid for retention in the final table. This approach assumes that an MS feature assigned to multiple samples contributes to the overall abundance value of the analyte in the pooled sample, thus carrying weight in the assignment. By analyzing intensity values in pool instances and the assignment locations, we can deduce which samples indeed contain an MS feature, eliminating the need for reanalysis by UPLC-MS.

The MMS-Q workflow builds upon the MMS computational deconvolution, using the presence/absence data as a blueprint of analyte assignment. The initial step in the MMS-Q workflow involves mapping the presence/absence data of each MS feature to the assigned sample location in the initial and rearranged grid pair(s) (**Figure 3.1.B**). For each mapped MS feature, the corresponding abundance value is appended to the appropriate row or column pool within the grids. Suppose the MS feature is not assigned to a sample in a row or column within the grid pairs following computational deconvolution but has a corresponding pooled intensity value. In that case, the pool intensity value is assigned as zero. The absence of a MS feature in a row or column occurs when the deconvolution algorithm does not assign a feature to a sample because it did not meet the requirements of being present in the sample in both grids and as a result, it is not retained. This process provides a visual representation of each identified MS feature within a grid pair, along with the intensity value in the pooled sample. The grid serves as a set of linear equations, where each sample is a new variable. Getting solutions for these equations was not always possible, especially when an MS feature appeared in multiple samples within a grid. Therefore, it was necessary to calculate at least an approximate solution for each identified MS feature in every sample.

The system of linear equations can be expressed as a matrix equation $Ax = b$, where A represents the matrix, with each column in the matrix corresponding to a sample in the sample set. The vector x contains unknowns, representing the relative abundance value of each MS feature in assigned samples. Vector b consists of constants that denote the detected abundance values of the MS feature in the row and column pooled samples. The matrix is depicted as an m x n matrix, where m represents the number of equations (pooled samples), and n indicates the number of unknown variables (samples within the grid). The values of x constitute column vectors of unknowns, representing the relative abundance values of the MS feature in each sample (**Figure 3.1.B**).

The matrix equation postulates that a possible solution vector x exists where b fits within the column space of matrix A. However, if A is not square ($m \neq n$), meaning there are more samples assigned to the grid than there are pooled samples ($m < n$), techniques like least squares are employed to find the best approximate solution $\hat{x}$ to solve the equation $Ax = b$. The least squares approach minimizes the sum of the squared differences between the entries of $A\hat{x}$ and b, where the solution vector $\hat{x}$ represents the values that minimize this difference.

The least squares method provides a solution for approximating the value of an MS feature within a pooled sample. This approximation can be visually represented by mapping it back to the binarized matrix. In cases where an MS feature is highly abundant in specific pooled samples but less prevalent in others, the weight assignment for that feature will be higher at the intersections of the more abundant pooled samples and lower in the less abundant ones (**Figure 3.1**). The weighted calculation will distinguish higher abundance MS features over less abundant assignments within the dataset. Such analysis empowers users to deprioritize or remove these assignments confidently during downstream analysis and instead focus on true positive MS features that contribute to the tested biological phenotype. Additionally, the MMS-Q workflow adds to the MMS method. Users can easily reanalyze legacy MMS data to add a relative abundance of information without recollecting MS data.

The following sections delineate a series of experiments designed to evaluate the efficacy of the MMS-Q workflow. These experiments encompass virtual multiplexing trials to assess the similarity between a predicted deconvoluted grid and ground-truth (GT) data. Additionally, a set of experiments investigates the method's capability to predict relative quantitation information from mixtures. Finally, the experiments conclude with an exploration of the potential of MMS-Q to assist in the prioritization of bioactive NPs following the reanalysis of a previous multiplexing experiment.

3.2.2. Simulation-Based Cosine Similarity Evaluation of the MultiplexMS-Q Method

Simulating sample pooling, a practical and efficient approach, provides a rapid means to gauge algorithmic success and evaluate the effectiveness of methodological adjustments. This includes subjecting the algorithm to rigorous stress tests, including

scenarios with a high presence of MS features, to ascertain the robustness in handling computationally challenging environments. Within MMS-Q, these challenging scenarios often arise when deconvoluting and assigning ubiquitous MS features to the correct samples. This also involves assessing the ability of the algorithm to predict the relative abundance of MS features within each sample based on the pooled sample MS data.

As highlighted in **Section 2.2.2**, identical MS features recurring across multiple samples within a grid pose a challenge in the MMSO sample strategy and computational deconvolution. This challenge can lead to incorrectly assigning detected features to samples, resulting in inaccurate relative abundance values of the assigned MS features. This issue stems from the understanding that the same MS feature, if present in multiple samples within a pool, contributes to the overall abundance value observed in the MS data.[119] The following details the development of a simulated pooling and computational deconvolution scheme, including the simulation of diverse matrices, the pooling of intensity values, and the application of the least squares algorithm for computational deconvolution. The overarching aim of this study is to meticulously evaluate the precision of analyte assignment, and the reliability of abundance value retrieval compared to a GT sample set.

The *in silico* algorithm employed here follows the MMS-Q workflow, including binarized computational deconvolution of pooled samples followed by relative abundance data retrieval. In contrast to the pipeline outlined in **Section 2.2.2**, the *in silico* process described here generates theoretical values within a defined range, reflective of the potential intensity values observed in a mass spectrometer and provided to researchers following MS file preprocessing. The algorithm carefully controls the number of features detected in a grid at each iteration. This means the algorithm generates data ranging from $x = 1$ to x = grid limitations (r^2), where x represents the number of occurrences in a grid. To simulate the pipeline, a sample list of size n is inputted into the algorithm. MMS-Q then generates symmetrical initial and rearranged grids of dimensions r x r for the simulation. In this experiment, grid sizes of r = 10, 20, and 30 were chosen, representing sample list sizes of n = 100, 400, and 900, respectively.

The generated initial and rearranged grids for this experiment start as empty matrices, occupied only by the coordinate positions of sample names slated for pooling.

An intensity value is then randomly generated within the range of 1E4 and 8E6 for each occurrence frequency x, ranging from $x = 1$ to $x = r^2$. The value is inserted in a random location within the initial grid. The assigned value range represents a common detectable intensity spectrum of MS features observed in untargeted metabolomics analyses. The algorithm then follows the MMSO protocol and generates a rearranged grid from the initial grid, ensuring that each x is assigned to the same sample between grids. These initial and rearranged grids serve as the GT for the MS feature location and quantitation in subsequent analyses. Each grid pair is processed through the MMS-Q pipeline i times, with each iteration containing different locations of x MS features within the grids. For this experiment, the pipeline was iterated $i = 400$ times for each occurrence across the three grid sizes, r. This means the pipeline went through 400 iterations for each occurrence x from 1 to r^2, using different sample assignments and abundance values.

Once values are placed in the grid, the rows and columns are separately summed (additive contribution of each occurrence x), generating a value corresponding to the overall intensity value of the detected MS feature(s) in a sample pool. This process is carried out for the initial and rearranged grids, creating a pool abundance value for each row and column. The pooled samples then undergo processing through the MMS-Q pipeline. Using the matrix equation, the least squares value for x is predicted by fitting the b values into the column space of A. This generates predicted intensity values for each retained MS feature following computational deconvolution. Utilizing the minimum intensity threshold feature, assigned MS features below the lower limit (1E4) are removed from the dataset.

The predicted values are organized into a list, with positions corresponding to the sample names. The predicted values from MMS-Q analysis are then compared to the GT values using cosine similarity scoring, generating values between 0 and 1. The cosine scoring method enables qualitative assessment of the algorithm's predictive capability in accurately generating values for each detected MS feature, even with an increasing frequency of MS feature occurrence. Briefly, cosine similarity is a measure to assess how similar two vectors are to one another, regardless of magnitude. The cosine similarity score is determined by the angle between the two vectors, ranging from -1 (diametrically opposed) and 1 (vectors are identical).[120] The lists are then binarized by

assigning a 1 to any non-zero number and used to calculate the false positive rate (FPR) (**see Section 3.3.4**).

The average cosine scores and FPR were plotted against one another, showing the degree of GT versus predicted value similarity as grid complexity increases (**Figure 3.2**). When using a $r = 10$ ($n = 100$) sized grid, the cosine similarity of the GT versus predicted is > 0.8 even when $x = 10$, a strong result considering there are 20 pooled samples per initial and rearranged grid. Also, the FPR at $x = 10$ is only 0.18 before jumping to 0.50 when $x = 20$. Increasing the grid sizes is expected to improve the cosine scores and FPR. An $r = 20$ ($n = 400$) grid size improves the cosine score at $x = 10$ to 0.9 while maintaining a low false positive rate at this same point (FPR = 0.05). Finally, the $r = 30$ ($n = 900$) grid size yielded a cosine score > 0.8 even when $x = 18$, corresponding to an FPR of 0.05. The results show that the algorithm has limitations when dealing with high MS features in a sample set and the accuracy of MMS-Q abundance value determinations after a multiplexing experiment. Since these data are generated using computer simulation, the results can be universally applied as guidelines for determining grid sizes when dealing with different sample complexities in MS data. These findings establish a benchmark for what can be expected in real-world laboratory experiments using complex sample datasets.

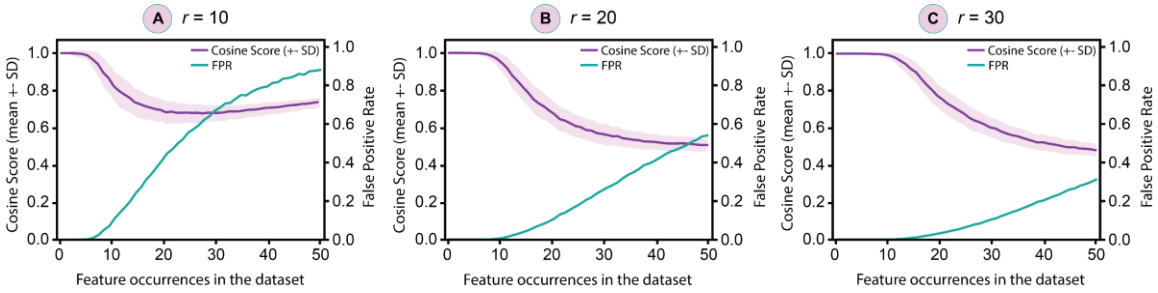


Figure 3.2 ***In Silico* Testing of the MultiplexMS-Q Strategy.**
The cosine score and FPR calculation results as the presence of an MS feature increases in a multiplexing experiment. The calculation has two purposes: assessing the similarity between predicted and GT values and determining the FPR after eliminating features below a minimum intensity threshold. The experiment was performed on three grid sizes: (**A**) $r = 10$, (**B**) $r = 20$, (**C**) $r = 30$.

3.2.3. Assessing Relative Abundance Retrieval in Multiplexed “One-Compound-One-Well” Scenarios

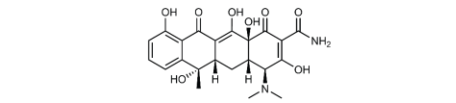
The performance of the MMS-Q algorithm is dependent on sample complexity and the frequency of identical MS features across samples. The *in silico* experiment in **Section 3.2.2** adeptly illustrates how sample complexity directly impacts the predictive capabilities of the MMS-Q algorithm, particularly when erroneous positive assignments have a discernable effect on the predicted abundance values. This was evaluated by comparing GT MS feature values against computationally derived values by cosine similarity scoring (**Figure 3.2**). The *in silico* experiment demonstrated the applicability of the MMS-Q abundance value predictions for increasingly complex grid deconvolution schemes. The MMS-Q was then applied to real-world experimental scenarios, evaluating algorithm performance with biological samples and the impact on MS data quality.

To evaluate MS feature assignment precision, a "one-compound-one-well" scenario utilizing a 5 x 5 MMSO strategy from the original MMS experiment was reused (**Appendix Figure A.4**). This endeavor aimed to gauge the MMS-Q precision in predicting intensity values for 25 laboratory standards. By simulating an environment where each well theoretically contains only one instance of a detected molecule, this experiment highlights the applicability and efficacy of the MMSO method in a simplified setting. The “one-compound-one-well” experiment serves as a “proof-of-concept” scenario, providing a benchmarking experiment to observe the potential effects of sample pooling, such as chromatographic overlap and ion suppression. Since these data were already acquired in the initial MMS study, the deconvoluted data underwent processing through the MMS-Q pipeline to generate predicted intensity values for each standard. The 25 standards underwent replicate analysis using an in-house Python script. If an MS feature is present in at least 2 out of 3 replicates for each set of triplicates, it is retained in the GT dataset, and the intensity values are averaged.

As an initial evaluation of the algorithm, the deconvoluted dataset was studied to assess the method's accuracy in assigning MS features to the correct sample. Consistent with the initial study's findings, all standards were correctly assigned in 25 of 25 cases, with two instances of false positive assignments demonstrating a precision rate of 96%.[109] Subsequently, the focus shifted to assessing the fidelity of predicted intensity values against averaged replicate samples. This qualitative comparison was

conducted one-on-one, evaluating predictive accuracy against the GT sample. Initially, each replicate set of laboratory standards were scrutinized for deviations between injections. The average standard deviation between replicate standards was 8.2% (**Figure 3.3, Appendix Table A.1**). Due to the matrix format and varying weights between samples where MS features are present, pooled replicate deviation values could not be calculated. Instead, examining the extracted ion chromatograms (EICs) of parent ions of each standard in assigned pooled samples revealed an average standard deviation of 9.7% from integrated values. The disparity in abundance values between computationally deconvoluted values and GT samples, including variations among all fragment values, was a mere 15.5% of the GT values. Furthermore, on average, the percentage difference between identified parent ions and the GT and calculated quantitative values was 9.8%, ranging from 0.2% (amoxicillin) to 35% (midecamycin).

Another aspect examined was the detection and abundance of associative fragment ions. Fragment ions associated with standards (presence and abundance values) were identified based on groupings in Progenesis QI analysis. These data were compared to the determined quantitative dataset. Overall, 92.7% of the fragment ions in the GT dataset were detected in the MMS-Q dataset. The success of this experiment underscores the predictive effectiveness and warrants further exploration into more complex scenarios to evaluate the predictive power of the MMS-Q algorithm when faced with realistic and intricate mixtures.



fragment *m/z*	ground truth			predicted
139.0563	7251.2640	±	558	8276.7546
152.0640	20784.0551	±	1625	23819.4742
154.0528	2573.7401	±	200	2837.6128
158.0980	9424.5164	±	1596	9524.1686
164.0642	10878.6113	±	781	12178.1923
176.0637	7065.2573	±	453	7770.3217
180.059	8329.0271	±	677	8924.9754
195.0830	3127.2913	±	811	2962.0573
203.0749	33093.9670	±	9262	30998.9568
254.0577	3388.8704	±	504	3321.9526
259.0627	5075.7953	±	871	5097.4654
280.0750	2839.2117	±	334	3109.2124
308.0690	2817.9551	±	177	3157.9563
382.0728	5230.6726	±	267	5973.0744
444.4238	35582.5718	±	2993	38988.8604
445.2610	20683.3799	±	5375	22929.4631

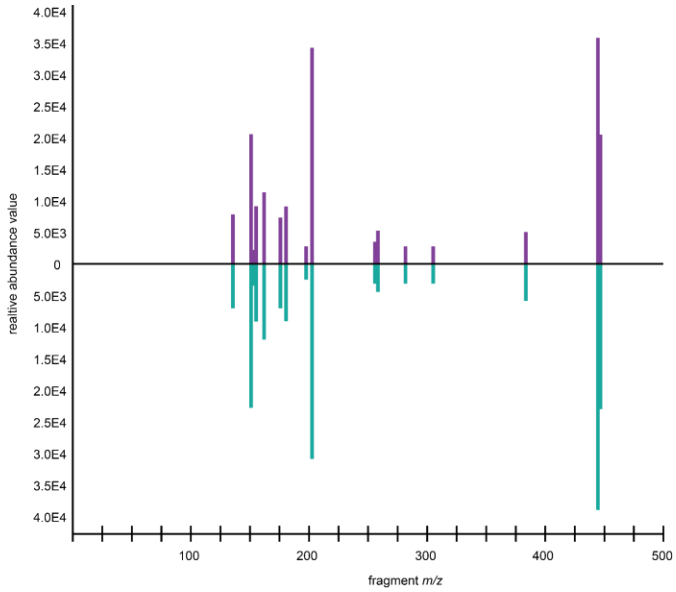


Figure 3.3 Comparison of Ground-Truth Tetracycline and Predicted Values from Pooled Samples.
A cross-comparison between a replicate compared tetracycline standard and relative abundance values determined from pooled samples. Sixteen total ions were associated with the parent ion based on MS data preprocessing and computational deconvolution. The average difference between the abundance values of tetracycline GT and the calculated values was 8.2%. All fragment ions were detected in both, which served as a benchmark experiment to assess the effectiveness of the pooling strategy in a “one-compound-one-well” scenario.

3.2.4. Evaluation of Single Replicate Standards Versus Standards Introduced into Complex Mixtures

The computational simulation of the MMS-Q pipeline demonstrated the high accuracy of the relative quantitation platform in assigning abundance values to MS features, mainly when the detection of MS features is low ($x < 5$). This observation was further supported by the 5 x 5 experiment, which mimicked a "one-compound-one-well" scenario. However, real-world metabolomics experiments rarely follow such ideal circumstances. These experiments typically involve complex mixtures containing a multitude of analytes, leading to background complexity and potential disruptions in peak shapes due to simultaneous analyte elution.

In the original MMS study described in **Section 2.2.5**, an experiment was devised to evaluate MMS performance using laboratory standards spiked into complex mixtures. This experiment included 90 bacterial fractionated extracts and ten commercially available NP, and semi-synthetic standards arranged in dual initial and

rearranged grids (**Appendix Figure A.8**). The experiment was designed to assess how effectively the deconvolution platform detected the ten standards in complex mixtures. In this section, we replicated the same experiment with a different focus: evaluate the accuracy of calculated quantitative values of the standards in the presence of complex backgrounds.

The commercial standards underwent quadruplicate analysis using interquartile range (IQR) replicate analysis, which eliminates outlier values to generate a GT dataset for comparison with pooled datasets (**Figure 3.4.A, Appendix Table A.2**). Comparison between the GT values generated from Progenesis QI and the MMS-Q calculated values provided insight into the accuracy of the algorithm in complex background scenarios. The MMS-Q calculated abundance values of the standards in the pooled samples differed from the GT dataset by an average of 8.4%, ranging from 1.8% (Novobiocin: GT – 34,400.93, MMS-Q – 34,802.67) to 25% (Azithromycin: GT – 12,356.84, MMS-Q – 15,428.54) (**Figure 3.4.E, Appendix Table A.2**). An outlier in the dataset was the prediction of the deferoxamine abundance value (GT – 1,866, Q – 24,978). This discrepancy prompted a reexamination of the pooled samples containing deferoxamine in Progenesis QI. The precursor $[M+H]^+$ *m/z* 561.3597 was identified in the assigned pooled samples but showed higher abundance values than the averaged GT value (55,336; 49,336; 43,347; 1,243). The high abundance values in the pooled data are likely due to a pipetting error contributing to a higher concentration of deferoxamine in the pooled samples. Despite this outlier, the correct sample assignment in 9 out of 10 samples provided sufficient evidence to proceed with the analysis.

A notable concern with sample pooling is the increase in inaccurate MS feature assignments, leading to high FPRs. One consequence of using the binarized system is obtaining an elevated FPR without the ability to ascertain the reality of these features without reexamining the raw MS data. The original MMS analysis demonstrated the ability to eliminate high frequency deconvoluted MS features without adversely affecting downstream bioactive compound discovery, as these features likely do not contribute to potential bioactivity (**see Section 2.2.6**). MMS-Q calculated abundance data enables analysts to address ubiquitous features with low intensities and prioritize less frequent molecules with higher intensities in the dataset that likely contribute to observed bioactivity. To simplify the dataset, users can set a predetermined minimum intensity threshold in the MMS-Q app by calculating the frequency of MS features within specific

abundance ranges. Setting the intensity cutoff threshold at the inflection point will remove MS features with low intensities (e.g., background ions, contaminants) to enhance analyte assignment precision before computational deconvolution.

A benchmark analysis of MMS-Q was conducted to determine the correct assignment of MS features to samples, particularly to assess whether standard features were assigned accurately. Encouragingly, all standard features were correctly assigned to the appropriate sample in all ten standards. This highlights the advantages of using MMS-Q in obtaining accurate abundance values despite complex backgrounds, further emphasizing the potential to identify bioactive molecules in mixtures accurately.

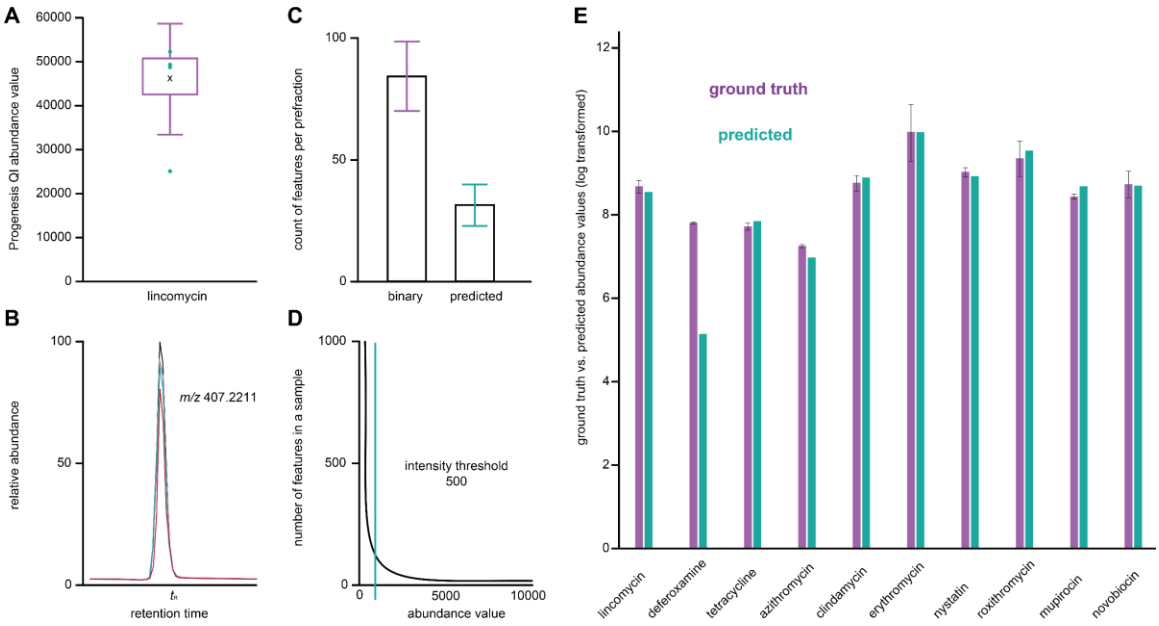


Figure 3.4 **Comparison of Ground-Truth and Predicted Abundance Values.** (**A**) Comparison of abundance values using IQR for lincomycin GT replicate standards. (**B**) EICs of $[M+H]^+$ *m/z* 407.2211 for lincomycin replicates, comparing intensity and peak shape. (**C**) MS feature count per prefraction using the binary MMS system and the MMS-Q predicted values (with minimum intensity threshold feature). (**D**) Determining minimum intensity threshold: plotting the frequency of intensity values for MS features against detected intensities. The minimum intensity threshold is determined by the point of inflection, below which the MS features are removed from the table. (**E**) Comparison of log-transformed abundance values between GT replicates and predicted values.

3.2.5. Leveraging Relative Quantitation for Accurate Detection of Bioactive Molecules

Understanding the relative abundance of an analyte across numerous samples is crucial for understanding biological processes, disease mechanisms, and responses to drug administration.[121,122] Initially, MMS analysis was designed to integrate rapid, high-quality UPLC-MS analysis of complex mixtures with concurrent bioactivity analysis of individual samples, mainly using CAM software like NP Analyst. The resultant MMS output file is binary, indicating only the presence or absence of analytes in the MS output. MMS-Q elevates this capability by affixing relative quantitative information to distinct MS features, facilitating more precise prioritization of identified analytes with relative abundance information.

In our prior study outlined in **Section 2.2.6**, we demonstrated the efficacy of MMS to screen large NP libraries using the sample pooling strategy. This was achieved by consolidating 925 fractionated microbial extracts into a 31 x 31 multiplexing experiment. The fractions were organized into a 31 x 31 grid, the smallest symmetrical grid size accommodating all 925 samples. Rows and columns of the grids were separately pooled and subjected to UPLC-MS as pooled fractions. This consolidation reduced the number of samples requiring MS analysis from 2,775 triplicate analyses to just 124, dropping the analysis time from weeks to a single overnight run. The results agreed with the original triplicate analysis, which identified several bioactive compounds from the NP Analyst network.[27,109]

Previous experiments have demonstrated that MMS is prone to a high FPR when pooling more than 20 samples together (**Figure 2.5**).[109] Additionally, the binarized data lack relative abundance information, leading to the prioritization of MS features based solely on Activity and Cluster scores without considering analyte abundance. These features may be minimal or absent in a fraction, necessitating reexamination of MS data for confirmation or rescreening.

Exploring the potential of MMS-Q, relative abundance values were calculated from pooled MS data using the original NP Analyst and MMS datasets, which included pooled MS information and BioMAP activity readings of individual fractions. The aim was to assess the accuracy of quantitation following computational deconvolution and to determine if compound prioritization of bioactive metabolites was more focused with a

minimum intensity threshold application (**Figure 3.5**).[27,109] The multiplexed MS data were processed via the MMS-Q pipeline to generate relative quantitative data for each detected MS feature. A minimum intensity threshold analysis was performed to remove MS features below the inflection point of intensity values (**Figure 3.5.B**). The MMS-Q metabolomics dataset and the initial BioMAP bioactivity data were provided to the NP Analyst platform for bioactive compound prioritization.

Initial observations showed a more focused network (**Figure 3.5.A**) of MS feature-prefraction relationships compared to the previous MMS analysis (**Figure 2.8**). The resulting network produced 21 distinct communities, with 136 possible targets above an Activity and Cluster score threshold of 2 and 0.3, respectively. Using the metabolomics dataset from the original MMS study, the same bioactive molecules were identified: micromonolactam (**3.1**), dracolactam A (**3.2**), amychelin C (**3.3**), collismycin A (**3.4**) and B (**3.5**), and SF2738D (**3.6**).[27,109] The calculated abundance values of these molecules were compared to values detected in adjacent prefractions (e.g., prefraction A vs. prefraction B vs. prefraction C), providing relative quantitation information of analyte presence across samples (**Figure 3.6**). This information will allow users to prioritize fractions for downstream isolation, assess the relative distribution of a molecule across a sample set, and eliminate the need for reanalysis of MS data for quantitative information.

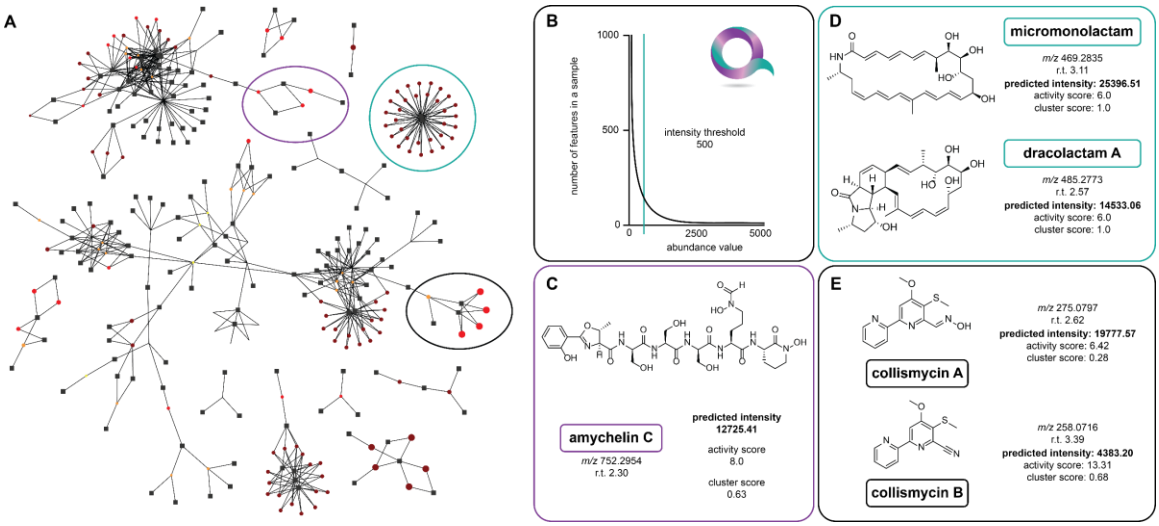


Figure 3.5 **Predicting Relative Abundance Values in Pooled Complex Mixtures to Identify Target Molecules Accurately.**
(**A**) A reduced NP Analyst output from MMS-Q with a minimum intensity threshold of 500 applied, increasing confidence in lead prioritization and minimizing false positives for a more targeted analysis. (**B**) Determining the minimum intensity threshold by plotting the frequency of the abundance values. For this experiment, a minimum intensity threshold of 500 was applied based on

the inflection point location. (**C**, **D**, **E**) The identification of the original bioactive molecules and determination of relative intensity values.

The results of the NP Analyst network (**Figure 3.5.A**) showed that all previously identified bioactive molecules (**3.1** – **3.6**) were detected in the benchmark experiment. The reduction in possible false positives and low abundance MS features were significant after applying a minimum intensity threshold based on calculated abundance values. Additionally, Activity and Cluster scores were compared to the original analyses to determine score variations using the new MMS-Q methodology (**Figure 3.5.C, D, E**). For instance, collismycin A (**3.4**, *m/z* 275.0797, t_R 2.62 min) had reduced Activity and Cluster scores (6.42 and 0.28, respectively) compared with the original analysis (**Figure 3.5.E**). This was due to the identification of collismycin A in an inactive fraction RLUS-2110A that fell just above the minimum intensity threshold (502), reducing the molecule's Activity and Cluster scores. After reexamination of the individual prefraction sample by MS, no presence of **3.4** could be detected, indicating a false positive assignment. In the original MMS-NP Analyst analysis, this precursor MS feature was not identified as potentially active by the NP Analyst program. This was due to the inaccurate assignments of the MS feature to multiple inactive fractions, lowering the Activity and Cluster scores below the cutoff values and rendering the feature inactive. However, the NP Analyst analysis flagged fragments associated with collismycin A (**3.4**) (e.g., *m/z* 205.0935) as active, with Activity and Cluster scores of 13.31 and 0.68, respectively. These scores prioritize this molecule for isolation and subsequent bioactivity screening. This experiment demonstrates how relative quantitation information can enhance the accurate prioritization of molecules within an NP Analyst network and help identify potential false positives after computational deconvolution in MMS-Q.

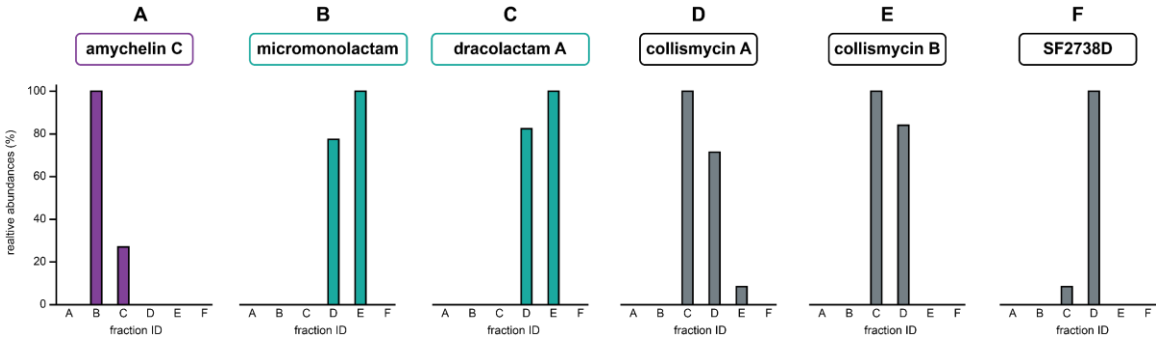


Figure 3.6 Relative Abundance Values of Bioactive Molecules Determined from MMS-Q.
(**A**) amychelin C (**3.3**) – fraction B: 12725.41, fraction C: 3435.66. (**B**) micromonolactam (**3.1**) – fraction D: 19788.51, fraction E: 25396.44. (**C**) dracolactam A (**3.2**) – fraction D: 11786.44, fraction E: 14533.06. (**D**) collismycin A (**3.4**) – fraction C: 19777.57, fraction D: 13702.69, fraction E: 1620.24. (**E**) collismycin B (**3.5**) – fraction C: 4383.20, fraction D: 3601.67. (**F**) SF2738D (**3.6**) – fraction C: 888.09, fraction D: 8536.49.

This analysis aimed to showcase the effectiveness of applying the MMS-Q methodology to HT analyses of complex mixtures. By calculating the relative quantitation of MS features across a dataset, researchers gain insights into prioritizing samples for isolating molecules and maximizing yields of isolated products. Integrating metabolomics data containing relative quantitation data with bioassay readings in CAM software platforms, such as NP Analyst, can provide valuable information on molecule behavior. For example, this approach allows monitoring activity versus abundance data to confirm the compound presence and may yield insights into molecule behavior in the bioassay (e.g., IC_{50} values). Overall, this demonstration highlights how MMS-Q can seamlessly incorporate into previous experiments and enhance the identification of potentially bioactive molecules in complex mixtures.

3.3. Conclusion

Technological advancements in HT metabolomics analyses have significantly accelerated the potential for discovering bioactive molecules from natural sources. A critical component in the discovery pipeline is the preprocessing software capabilities, which include advanced abilities to discern essential and statistically significant MS features even in complex mixtures where similar elution times can complicate peak shapes.[14] This increase in peak-picking sensitivity and accuracy enables the accurate

determination of relative abundance values between MS features across replicates in a dataset and throughout a sample set.

Despite these advancements, the HT UPLC-MS analysis of ever-expanding sample libraries hinders the discovery process due to chromatographic elution times.[19] In response to this bottleneck, MMS was developed to increase the throughput capabilities of UPLC-MS analyses by pooling multiple samples together and then computationally deconvoluting detected MS features to reconstruct individual samples after MS analysis. Specific experiments and rigorous algorithm testing outlined in **Chapter 2** demonstrated the success of this methodology in reducing the analysis time for an extensive, fractionated library using UPLC-MS. This characteristic sampling strategy improves throughput and maintains high-quality MS data. The binarized processing in the computational deconvolution algorithm contributes to the speed of MMS, where the output for the reconstructed sample lists indicates the presence and absence of data (1/0).

As shown in **Section 2.2.6**, the binarized output does not hinder the discovery of potentially bioactive metabolites in complex mixtures, particularly when using CAM platforms such as NP Analyst.[27] However, the binarized data prevents the relative abundance comparison of an MS feature distributed across a sample set without reanalyzing the sample. Determining the relative abundance of MS features in pooled samples presents a challenge and, to the best of our knowledge, has not been attempted before in metabolomics. The new addition, MMS-Q, calculates the relative abundance values of MS features from pooled row and column samples. These calculations are based on the binarized deconvoluted results from MMS. The deconvoluted data are formatted into a matrix equation for each grid pair. In this equation, each presence value contributes a weight to the overall pool abundance value. The relative abundances of each presence assignment within the grid are then calculated using least squares. The values between the initial and rearranged grid pairs are then averaged to obtain the final relative abundance value (see **Section 3.2.1**).

For several reasons, the relative abundance value assignment of MS features in samples is essential. It normalizes variations in sample size and total metabolite content, making comparisons across samples more meaningful.[123] Additionally, assigned values allow researchers to compare metabolite levels between different conditions when

performing metabolomics experiments (e.g., experimental versus control conditions).[124] Focusing on relative changes of a molecule across a sample set biologically enables the identification of metabolites associated with specific biological processes or pathways.[123,124] Finally, including relative abundance values enhances data visualization techniques, such as heatmaps and scatterplots, by adding an extra dimension for analyzing MS feature distribution.[125]

This chapter outlines the experiments and testing that went into the development of MMS-Q. These experiments demonstrate the accuracy of MMS-Q in determining the relative abundance value of an MS feature in a pooled sample. The MMS-Q workflow enhances the original MMS workflow by more accurately assigning relative abundance values to deconvoluted assignments. The workflow can be easily incorporated into the original MMS pipeline. Users can choose their analysis preference by toggling between the 'Deconvolute' and 'Quantitation' tabs in the graphical user interface (GUI) (**see Section 3.4.5**). Since MMS-Q leverages the output from MMS, it can seamlessly integrate into existing analyses. Integrating MMS-Q into the metabolomics analysis pipeline enables a more accurate and efficient analysis of complex sample sets. This enhancement reduces analysis time, providing researchers with critical relative abundance data. This facilitates deeper insights into the biological significance of metabolites.

3.4. Material and Methods

3.4.1. General Experimental Information

The solvents use in the MS acquisition were of optima LCMS grade. Laboratory antibiotic standards were employed without any additional purification steps. The MMS-Q pipeline and GUI were developed using the Python coding language.

3.4.2. UPLC-MS Conditions

The measurements were conducted using an ACQUITY *i*-Class UPLC system manufactured by Waters Corp. The UPLC system was equipped with an ACQUITY HSS T3 column (1.8 µm, 2.1 × 100 mm, Waters). A linear elution gradient was employed for separation (mobile phase A: H_2O + 0.01% formic acid; mobile phase B: acetonitrile +

0.01% formic acid, 0.5 mL/min) as follows: 0−0.3 min, 5% B; 0.3−4.7 min, 5−90% B; 4.7−5.5 min, 90−98% B; 5.5−5.8 min, 98% B; 5.8−7.5 min, 5% B. MS data were acquired using SYNAPT G2-Si qTOF (Waters). All mass measurements were recorded using ESI+ data-independent acquisition experiments.

3.4.3. Mass Spectrometry Data Processing

All samples were processed using the Progenesis QI software suite (v3.0.8293.38961, Nonlinear Dynamics, Waters). Lock-mass corrected positive data was uploaded into Progenesis QI for spectra alignment and peak picking using default settings. Positive polarity adducts were detected based on $[M+H]^+$, $[M+Na]^+$, $[M-H_2O+H]^+$, and $[M+K]^+$ ions. The sample set then underwent blank subtraction and minimal intensity threshold filtration. The generated feature table was rearranged to organize the data into a flat csv file, with MS features represented along one axis and pooled sample names along the other. The rearranged feature table was provided to the MMS-Q app for computational deconvolution and prediction of relative abundances.

3.4.4. *In Silico* Generation of Random Grids for Relative Abundance Prediction

The *in silico* subsampling approach was used to assess the accuracy of abundance values following computational deconvolution of pooled sample data. The algorithm is implemented in the Python coding language, utilizing libraries such as NumPy,[126] pandas,[127] and scikit-learn (sklearn).[128] Random values between the specified intensity limits are placed in a grid of a predetermined size, *r*. Using the MMSO protocol, a rearranged grid is generated from the initial grid (**Figure 2.3**). These initial and rearranged grids serve as the GT dataset for comparison.

The values in each row and column of the grid are summed to assign the abundance value of an analyte in a pooled sample. Pooled samples are then processed through the MMS-Q pipeline, generating abundance values for each sample. This cycle is repeated 400 times with new values and locations for each iteration. The generated abundance lists are compared to the GT values using cosine similarity:

$$K(X,Y) = \frac{<X,Y>}{(\|X\| * \|Y\|)} \qquad \textbf{(Equation 3.1)}$$

Where *X* is the list of GT values, and *Y* is the generated abundance values from MMS-Q. The quantity of MS features in the grid incrementally increases by $i + 1$ until $i = 50$, representing the presence of an MS feature in 50 different samples. Finally, the FPR calculation follows the outlined procedure in **Section 2.4.7**.

3.4.5. MultiplexMS-Q Application Development

MMS-Q builds upon the existing MMS infrastructure, utilizing the computational deconvolution algorithm to demultiplex samples and GUI for easy usability. Changes have been made to the GUI to accommodate the MMS-Q pipeline. Firstly, the app has four processing options: Preparation, Deconvolution, Quantitation, and Cleaning. Secondly, the data processing uses a sample directory that contains the MMSO grids, preparation tables, raw and processed files, and the results folder. This addition to the data processing will allow users to keep information organized as this data processing method can produce large datasets and prioritize essential data files for downstream analyses.

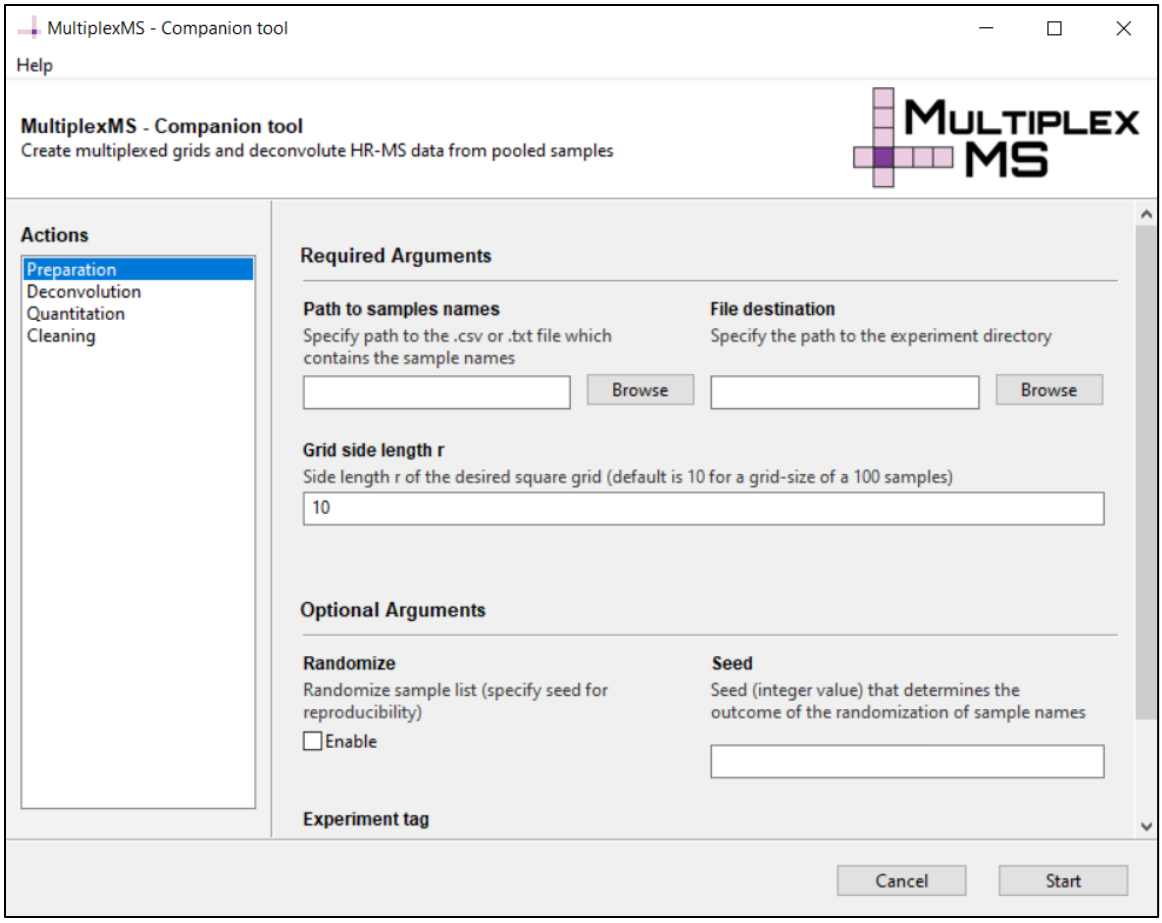


Figure 3.7 Screenshot of the Preparation Page of MultiplexMS-Q.

The experimental sample lists are provided to the app in the 'Preparation' tab and arranged into initial and rearranged grids per the user specifications. The specifics of this step are outlined in **Section 2.4.6**. Following the acquisition and preprocessing of data, the pooled feature table consisting of abundance data is provided back to the GUI and computationally deconvoluted under the 'Deconvolution' tab. This procedure produces a binarized demultiplexed file with presence and absence information for detected MS features. The deconvolution step in MMS-Q produces two files contained in the 'processed' folder of the experimental directory: 1) 'binary_deconvoluted_feature_table' which contains the presence and absence of an MS feature assigned to specific samples, and 2) 'binary_deconvoluted_feature_table_with_placeholders' containing the same presence and absence data, but contains the 'placeholder' samples if used to fill empty spaces in the symmetrical initial and rearranged grids.

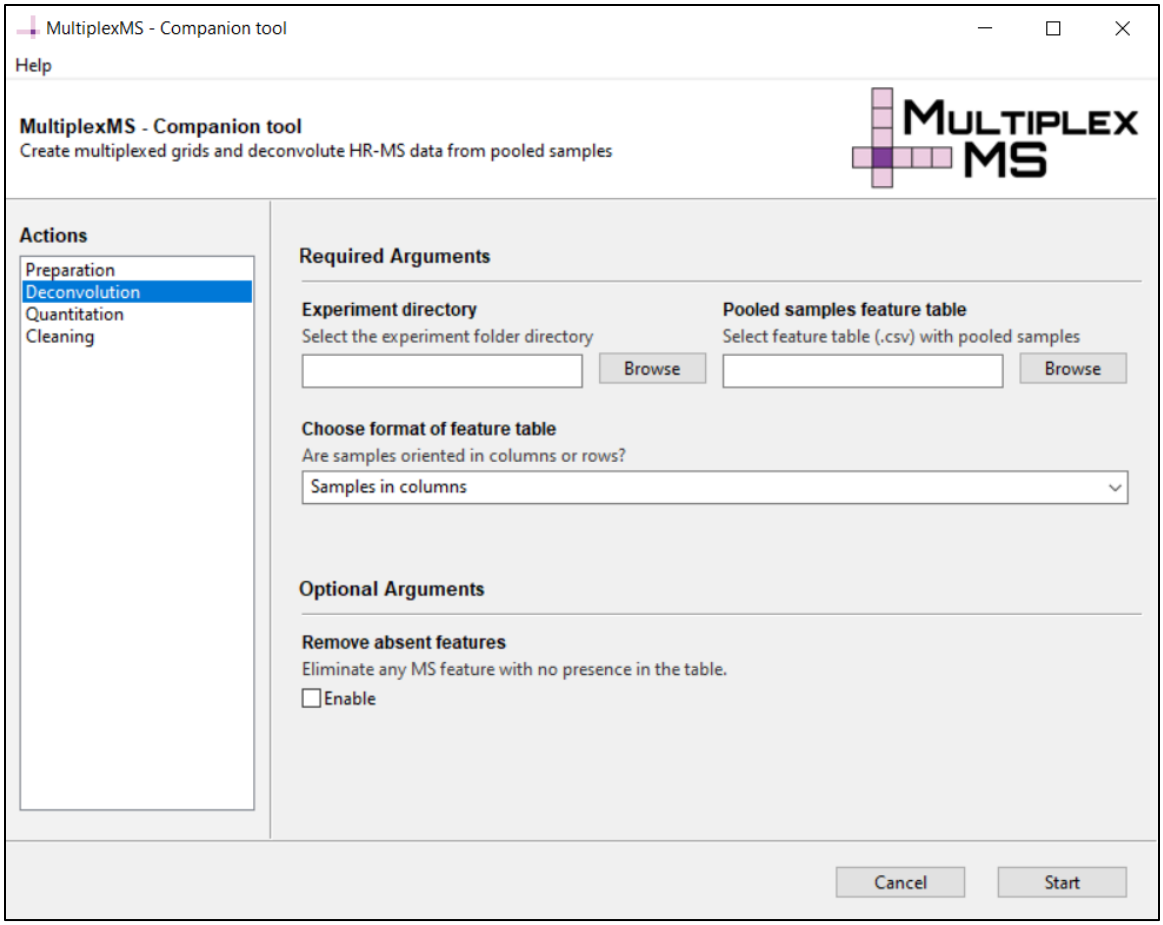


Figure 3.8 Screenshot of the Deconvolution Page of MultiplexMS-Q.

The 'Quantitation' tab is kept separate from 'Deconvolution' if users do not want to use the entire quantitation pipeline and only need the binarized data. The 'Quantitation' tab is simplified, allowing users to direct the app toward the experimental directory and letting the app do the processing automatically. The app uses the experimental directory to look for the 'processed' folder containing the 'binary_deconvoluted_feature_table_with_placeholders' file and the pooled feature table containing the abundance data. The MMS-Q algorithm uses the binarized deconvoluted data to assign features and calculate the relative abundance data. MMS-Q processes these data feature-by-feature, creating individual grids for each MS feature containing the binarized deconvoluted information with appended pooled abundance data corresponding to each row and column. These initial and rearranged grids are saved as separate csv files in their respective folders, 'initial_grid_decon_with_intensities' and 'rearranged_grid_decon_with_intensities,' where each file has a designation '(initial_or_rearranged)_grid_(grid_number)_row_and_column_intensities_(retention_tim

e)_(*m/z*_value).' The quantitation tool in MMS-Q uses these files to calculate the relative abundance values for each MS feature present in each grid using matrix equations and least squares calculations. Once a value is assigned in the initial and rearranged grid corresponding to a certain MS feature, the values are averaged between the grids, as described in **Section 3.2.1**. The averaged values for all MS features in each sample in each grid (grid_1 → grid_*z*) are arranged in a flat csv file and exported to a folder 'quantitative_grids.' Each file in this folder corresponds to an initial and rearranged grid in an experiment. Once all the quantitative grids are generated, all the files in the folder are concatenated into a table containing all the samples in the multiplexing experiment. This table containing the calculated relative abundance values of each detected MS feature in all samples is exported to the 'results' folder of the experimental directory. The relative abundance values depend on each assignment's weight within the grid and the corresponding pooled abundance value. Depending on the weight, values within the grid may correspond to low abundance or even negative values. Upon the user's discretion, these values can be removed from the grid using a determined minimum intensity threshold value. The Critical threshold option eliminates MS features in more than x% of samples. This option eliminates ubiquitous features from a dataset and prioritizes the rare molecules.

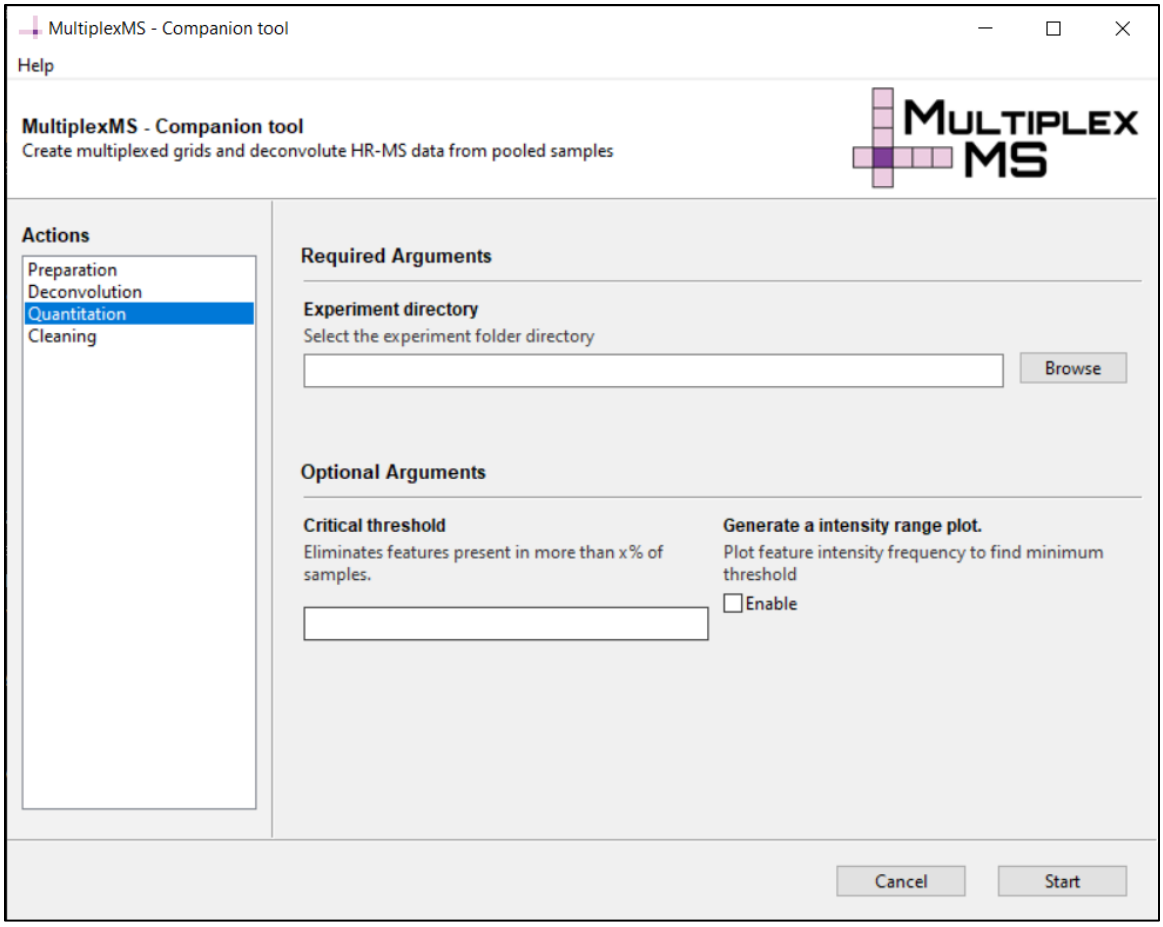


Figure 3.9 Screenshot of the Quantitation Page of MultiplexMS-Q.

A new addition to the existing MMS pipeline is the ability to determine a minimum intensity threshold cutoff value, regardless of any computational deconvolution. This value, recommended to be set at the inflection point of the intensity range plot, allows users to eliminate MS features in a pooled or deconvoluted sample below the cutoff value. This process helps improve the precision of assigned MS features and the accuracy of calculated abundance values. Also on this page is the 'Remove Absent Features' option, which eliminates any MS feature that is not present in any sample. This option is critical for CAM experiments using platforms such as NP Analyst, which require each MS feature to be present in at least one sample.

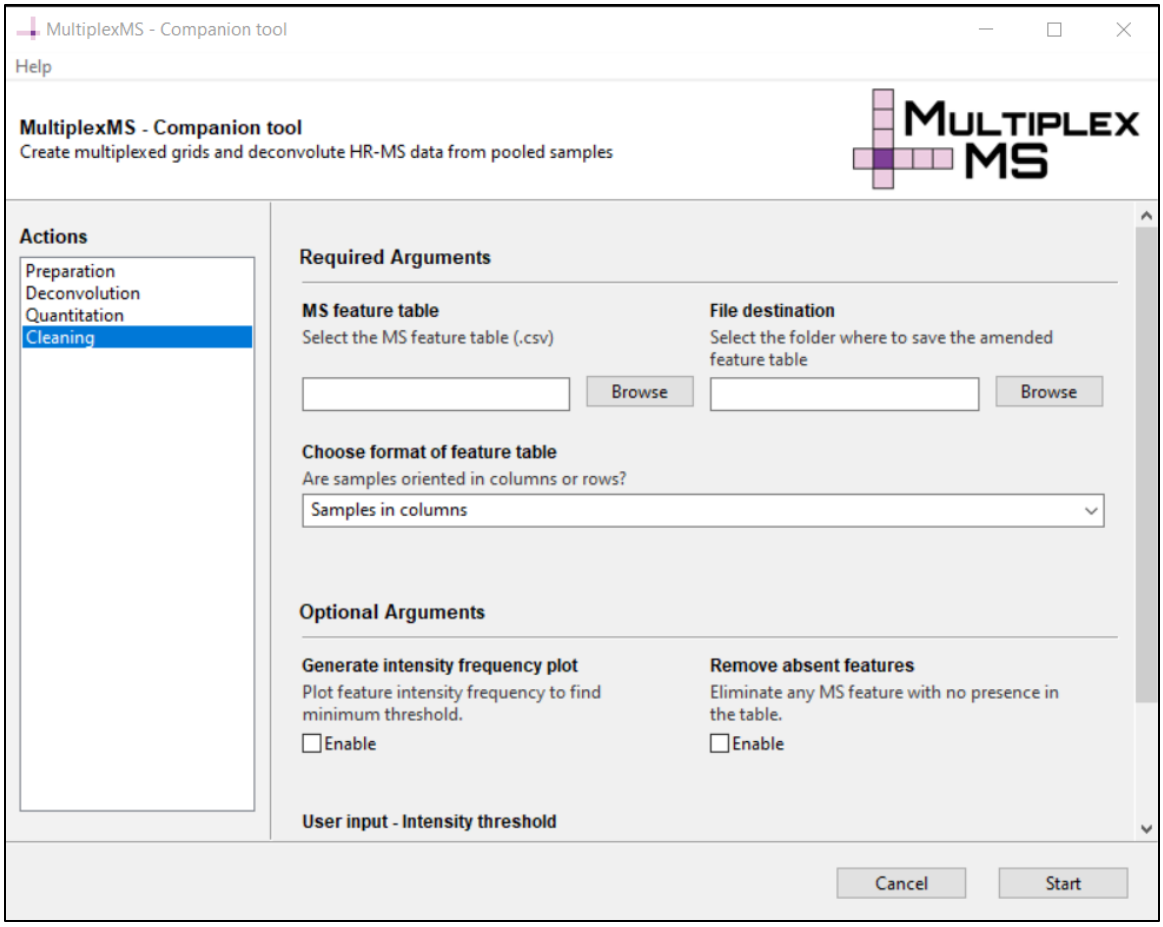


Figure 3.10 Screenshot of the Cleaning Page of MultiplexMS-Q.

Chapter 4.

High-Throughput Potential of Next-Gen Mass Spectrometers in Complex Mixture Feature Recovery

Authors: Michael J. J. Recchia, Trevor N. Clark, Adam King, Lee Gethings, David Heywood, Roger G. Linington

Author Contributions: M.J.J.R. and R.G.L. designed the study. M.J.J.R. performed the sample pooling and acquired the MS data on the SYNAPT G2-Si mass spectrometer. A.K. and L.G. collaborated closely to acquire the MS data on the Xevo MRT. M.J.J.R. processed the MS data from the two instruments and performed replicate comparison, blank subtraction, alignment, and retention time correction. D.H. assisted with experimental logistics and provided Progenesis QI software for data analysis. T.N.C. assisted with data acquisition and instrument parameter settings.

4.1. New Instrument Technologies in Bioactive Molecule Discovery

4.1.1. Introduction

In the rapidly evolving field of bioactive molecule discovery, it is crucial to stay at the forefront of technology to maintain a competitive edge. These technological advancements are occurring in an era of sophisticated high-throughput (HT) techniques, which allow researchers to identify biologically active molecules from vast libraries targeting specific biological mechanisms.[27,74,129] Indeed, having an advantage in this field helps accurately determine molecular properties and aids in discovering new structures. Innovative bioassays, such as Cell Painting, have opened new pathways for drug discovery, emphasizing the importance of accurately annotating molecules in HT screening campaigns of complex mixtures.[33]

Mass spectrometry (MS) has revolutionized drug development and analysis since its inception in research.[71,72] This ultra-sensitive technique measures qualitative and quantitative information about molecules in complex mixtures, making this an invaluable tool for compound characterization.[106] MS technology enables atomic-level analysis of

drug molecules, accelerating the identification of potential drug candidates.[71,130,131] MS also plays a critical role in the 'Omics' research fields, such as proteomics, which provides insights into disease mechanisms and potential therapeutic targets, and metabolomics, which helps understand metabolic networks in biological samples.[4,5,113,132–135] This advancement, propelled not only by instrument technology but by progressive experimental methods, preprocessing software algorithms, artificial intelligence, and publicly accessible data repositories, has moved omics fields beyond qualitative analyses to a range of global and targeted quantitative methods that continue to revolutionize drug discovery and development.[5,24,53,136–138]

MS technology has advanced in three main areas: resolution, sensitivity, and speed.[10] High-resolution (HR)-MS, the use of mass spectrometers with high resolving power capabilities (R > 20,000 full-width half maximum (FWHM) and mass accuracy < 5 ppm), has significantly improved mass accuracy measurements.[139] It allows the distinction between compounds with the same nominal mass, determines elemental compositions, and provides more precise chemical insights into complex samples.[10,133] HR-MS is crucial for omics analyses, particularly metabolomics, where accurate identification of potentially thousands of molecules is required.[5,133] HR-MS enhances our understanding of the chemical space by unambiguously assigning molecular formulae to various features in a dataset, especially when mass accuracy measurements are sub-1 parts per million (ppm).[53,140] The HR capacity can further assist in distinguishing similar molecules by resolving fine isotope peaks in the mass spectrum (**Figure 4.1**). Overall, leveraging powerful instrumentation enhances confidence in molecule identification and provides valuable insights into the composition of unknown metabolites.

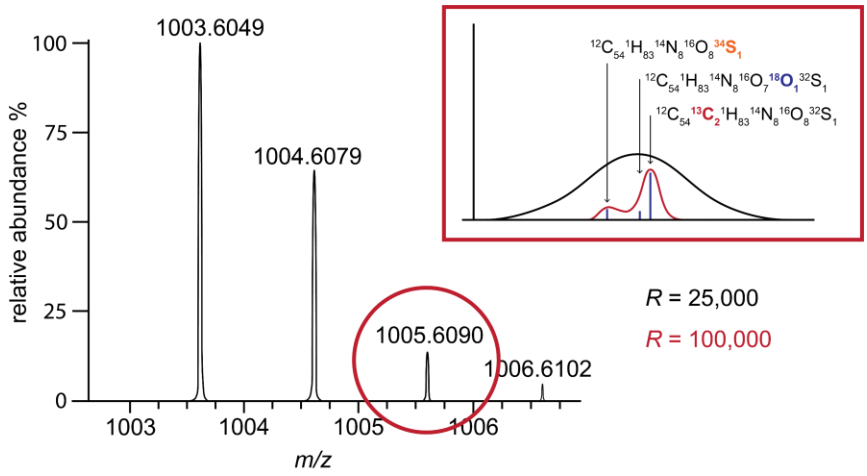


Figure 4.1 **Example of High-Resolution MS to Distinguish Fine Isotopes.** An example of HR-MS between instruments differing in resolving power. Instrument 1 has a resolving power of 25,000 FWHM, while Instrument 2 can obtain 100,000 FWHM. The advantage of using higher-resolving-power instruments is the identification of fine isotopes. Identifying these elements can increase confidence in molecular formula determinations and improve molecule identification.

4.1.2. High-Resolution Mass Spectrometry in Natural Products Metabolomics

Investigating the diversity of molecular formulae in natural products (NPs) and the ability of modern HR-MS instrumentation to distinguish between molecules is a significant advancement in metabolomics research. The NP Atlas compound repository, a database of over 36,000 reported microbial NPs, provides a platform to study molecular formula diversity and understand how well HR-MS instrumentation can potentially identify molecules in mixtures.[53] Morehouse *et al*. examined the occurrence, distribution, and grouping of microbial NPs to test the hypothesis that intra-family formula distributions are diagnostic identifiers for NP compound families.[129] A re-examination of the updated dataset identified 14,603 distinct molecular formulae consisting primarily of carbon, hydrogen, and oxygen elements, consistent with findings by Morehouse *et al.* (**Figure 4.2**). In their study, excluding formulae that occurred only once, the researchers found that molecular formulae were highly diagnostic for the compound family.[129] By leveraging the HR capabilities of MS instrumentation, we can discern a high distribution of molecular formulae in the reported microbial NP world, allowing the distinction between specific molecules, even those with similar nominal masses. This leads us to the intriguing question: Does nature provide a wide range of various structures of

elemental configurations to allow instrumentation to distinguish between molecules, or is there a large overlap in molecular formulae?

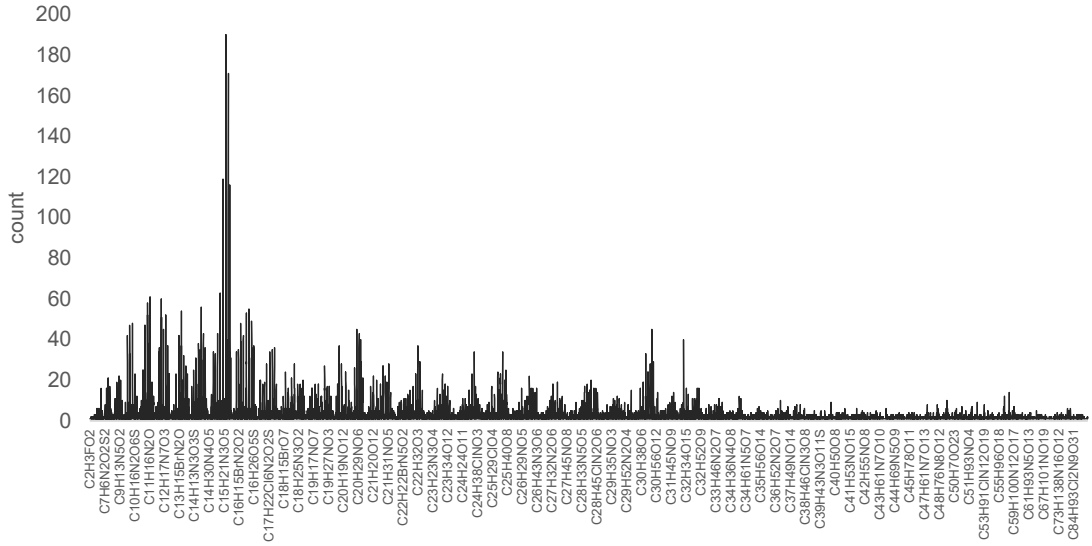


Figure 4.2 Distribution of All Reported Formulae in the NP Atlas.

Analysis of the reported NP Atlas library revealed 14,603 distinct molecule formulae. Among these, 9,425 formulae appeared only once, 2,325 appeared twice, and 931 appeared three times (**Figure 4.3**). Notably, 1,484 formulae contain at least one chlorine atom, while 283 contain at least one bromine atom, both exhibiting a characteristic isotope pattern in the compound mass spectrum. These analyses demonstrate the distribution of molecular formulae across the known NP chemical space and emphasize the practical advantages of using HR-MS to measure the mass of molecules to determine elemental composition accurately. With cutting-edge MS technology, mass measurements can be measured to 5 decimal places, allowing mass accuracy at the parts-per-billion (ppb) level. This precision facilitates the unequivocal determination of molecular formulae, aiding in the dereplication process to identify known or novel molecules and thereby contributing to the advancement of metabolomics research.

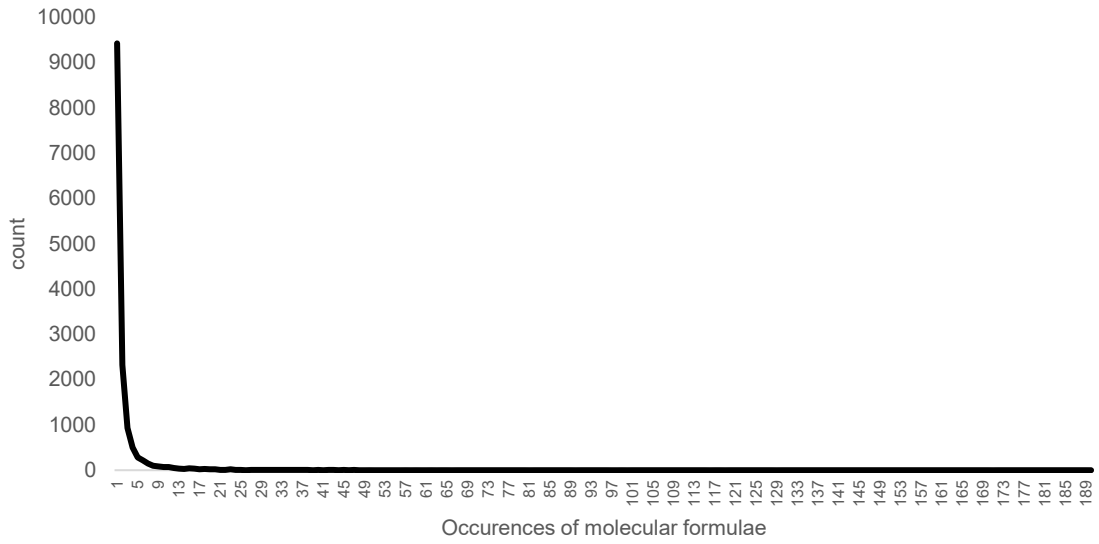


Figure 4.3 Frequency Distribution of Molecular Formulae in the NP Atlas.

4.1.3. Advances in Mass Spectrometry Instrumentation

Enhancements in ionization techniques such as matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) have improved the detection capabilities of MS analysis. These techniques allow for analyzing a wide range of molecules across diverse samples. Mass analyzers, such as quadrupole time-of-flight (qTOF), have reduced analysis times, making them valuable for large-scale screenings.[5,43,73–75,141,142] Coupling MS with ultra-performance liquid chromatography (UPLC) further refines the analysis of complex mixtures by chromatographically separating molecules before MS analysis.[5] UPLC-ESI-qTOF-MS (referred to as UPLC-MS in this chapter) has become essential in metabolomics for analyzing complex mixtures from extensive extract libraries.[5,37,73]

Despite these advancements, UPLC-MS analysis of large extract libraries has limitations. As discussed in **Chapters 2** and **3**, the time required for chromatographic separation before MS analysis can be a bottleneck in the discovery process. **Chapter 2** introduced a novel sampling strategy called MultiplexMS (MMS) to reduce the analysis time by pooling multiple samples for simultaneous UPLC-MS analysis. The complexity of samples and matrix effects can lead to ion suppression, affecting the precision of the demultiplexing methodology.[143] The previous chapters explored these concerns,

describing techniques and strategies to mitigate data loss caused by pooling complex samples.

The study described in **Section 2.2.4** illustrates the limitations of sample pooling using Waters UPLC in tandem with a Waters SYNAPT qTOF G2-Si equipped with an ESI ionization source – an industry-leading instrument for analyzing complex mixtures. The study examined a fractionated microbial extract containing known NPs in triplicate to establish a ground-truth (GT) dataset. Additional fractionated extracts were incrementally pooled until the number of samples reached $r = 100$. UPLC-MS was then used to analyze mixtures to assess the retention of GT features amid increasing sample complexity. The study found that 78% of GT features were retained when pooling 30 complex samples (**Figure 2.6**).[109]

Waters Corp. has developed a new qTOF technology called the Xevo™ Multi-Reflecting Time-of-Flight (MRT). This technology offers improved sensitivity, resolving power (100,000 FWHM), and acquisition speed (100 Hz), allowing it to obtain parts-per-billion (ppb) mass accuracy (**Figure 4.4**). In the study described in this chapter, the '1-to-100' experiment from **Section 2.2.4** was conducted in collaboration with Waters Corp. using the Xevo MRT, and the results were compared to those acquired on the SYNAPT G2-Si using the same dataset. The assessment involved a two-part evaluation: first, by identifying the three known molecules in the GT sample throughout the experiment, and second, by measuring the retention of GT MS features despite increasing sample complexity. It is important to note both instruments used identical ionization sources (ESI) and UPLC parameters, differing only in the MS technology. This chapter provides insights into how enhanced instrument resolution, sensitivity, and scanning contribute to improved feature recall and peak capacity. These findings serve as a benchmark for future multiplexing experiments and highlight the accuracy of feature retention achievable using HR-MS instruments.

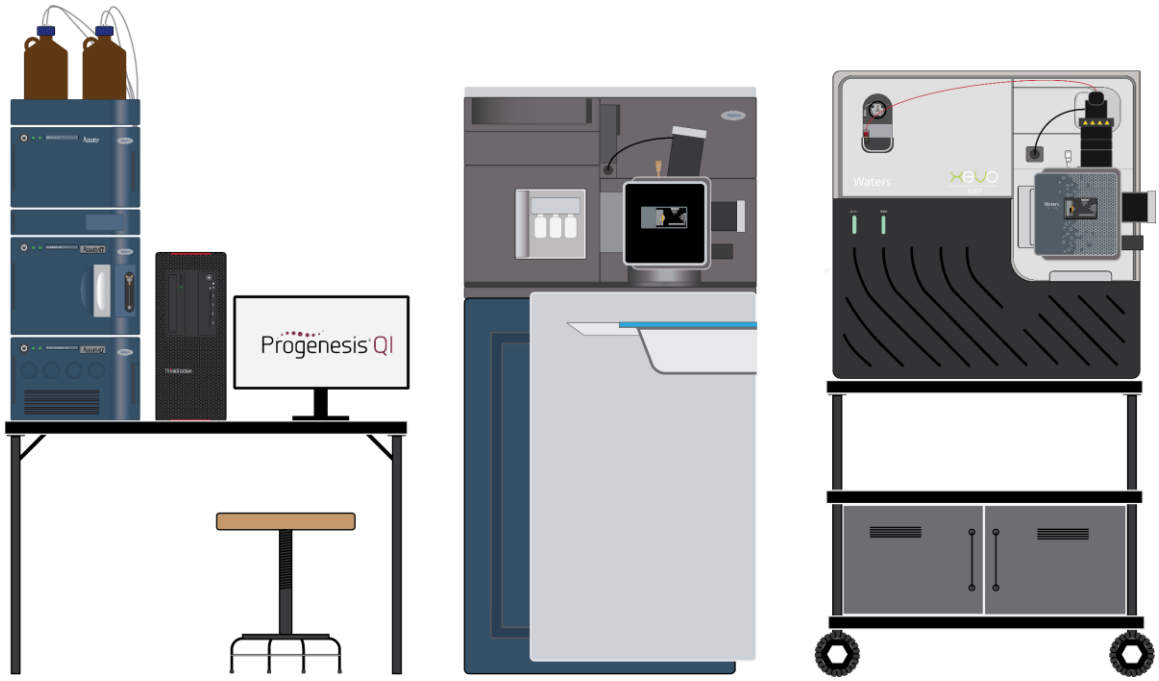


Figure 4.4 Instrumentation Used in Assessing the Advantages of Using HR Mass Spectrometers When Analyzing Complex Mixtures.
Waters HR-MS instrumentation was used in this experiment. A Waters UPLC *i*-class system was used for the chromatography on the front end of the two mass spectrometers. The first instrument was a Waters SYNAPT G2-Si with an ESI source in MSe resolution mode. This acquisition mode has a resolving power of 25,000 FWHM with a 5 Hz scan speed. The second instrument was the Xevo MRT. The same ESI source was used with this instrument and is capable of scan speeds of 100 Hz with a resolving power of 100,000 FWHM. MS data was processed using the Progenesis QI software suite.

4.2. Instruments and Software Used in this Study

4.2.1. Waters ACQUITY UPLC *i*-Class

A Waters UPLC *i*-class was used for front-end chromatographic separation before MS analysis for this analysis. The UPLC system has a binary solvent manager (BSM), a sample manager with a fixed loop (SM-FL), and a column heater. Chromatographic separation is achieved using an ACQUITY HSS T3 column - a fully encapped, low-coverage C_{18} bonded-phase column designed to retain extremely polar compounds using reversed-phased chromatography with high aqueous mobile phases. This instrument setup enhances resolution and ionization efficiency, resulting in narrow peak widths and reduced carryover between samples. For the chromatographic details, see **Section 4.5.2**.

4.2.2. Waters SYNAPT G2-Si (ESI-qTOF-MS)

The Waters SYNAPT G2-Si MS system combines StepWave™ ion optics with Quantitative Time-of-Flight (QuanTOF™) and High Definition Mass Spectrometry® technologies. The SYNAPT MS offers 25,000 FWHM resolution in 'resolution' mode with a <2 ppm mass error and a scan rate of 5 Hz in MSe resolution mode. StepWave technology reduces background noise and eliminates neutral compounds, achieving a higher signal-to-noise ratio than traditional qTOF instruments. The instrument features HR-MS capabilities and a dynamic range of up to 10^4. It has an ESI ionization source for robust data-independent acquisition with HR and mass accuracy. For details regarding UPLC-MS conditions, **see Section 4.5.2**.

4.2.3. Waters XEVO MRT (ESI-qTOF-MS)

The Waters Xevo™ MRT MS shares similar instrument components with the SYNAPT G2-Si MS, including the StepWaveXS™ ion guide technology for enhanced signal-to-noise. The Multi-Reflecting TOF (MRT) technology, which reflects ions 8x between two gridless electrostatic mirrors on a 4m flight path, significantly boosts resolving power and sensitivity. The Xevo MRT delivers an impressive 100,000 FWHM resolving power, with a mass accuracy of <500 ppb and acquisition rates up to 100 Hz. For details regarding UPLC-MS conditions, **see Section 4.5.2**.

4.2.4. Nonlinear Dynamics Progenesis QI (Waters)

The Nonlinear Dynamics MS processing software, Progenesis QI, enables users to accurately quantify and identify compounds to support omics research. This study used Progenesis QI to process the raw datasets acquired on both instruments. For efficient and objective analyses, raw MS files were imported into Progenesis QI, and a reference run was selected to align all runs in an experiment. The processing selects preferred adducts and peak picking to create an aggregated dataset from all aligned runs. Progenesis QI manages complex samples, discerning overlapping compound ions to detect complex datasets accurately. The processed data can then be used in downstream statistical analysis and compound identification. Since the raw MS data was acquired on two different instrument platforms, various processing steps are needed to

manage data interpretation; therefore, each dataset must be processed in separate Progenesis QI experiments (**Section 4.5.3**).

4.3. Results

4.3.1. Impact of Instrument Choice on Mass Spectrometry Data Quality in Complex Samples

An essential aspect of HT metabolomics is acquiring high-quality molecular data, especially as sample complexity increases. This complexity can compromise accurate mass measurements due to overlapping MS signals. Precise mass measurements enable researchers to obtain unequivocal molecular formulae, aiding in discovering new molecules and facilitating the dereplication of known compounds. This study aimed to assess how much an instrument influences the quality of MS data, particularly as the background matrix becomes more complex with the addition of intricate mixtures. The experiment outlined in **Section 2.2.4** was recreated, analyzing samples on a Waters Xevo MRT and a SYNAPT G2-Si MS instrument.

The ground-truth (GT) fraction from a microbial extract contains multiple molecules, three of which were previously identified as micromonolactam (**4.1**), dracolactam A (**4.2**), and dracolactam C (**4.3**) (**Figure 4.5**).[27] More fractions from various extracts were added while maintaining a constant concentration of the GT sample throughout the experiment. UPLC-MS was used to analyze the mixture at each addition step to determine if the instrument could detect the known GT molecules and if there was a detection limit due to sample complexity.

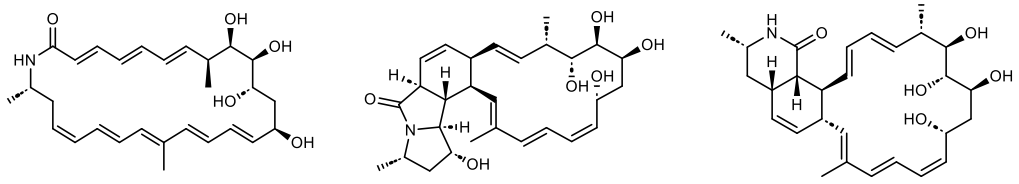


Figure 4.5 Target Molecules for the '1-to-100' Multiplexing Experiment.

After acquiring the dataset on the Xevo MRT and re-acquiring data on the SYNAPT G2-Si, raw MS data from both instruments were analyzed using Progenesis QI

for alignment and peak picking. The exported data was then compared for MS feature comparison between datasets (**see Section 4.5.3**). Compounds **4.1** – **4.3** were GT targets to check for consistent detection across all samples. However, dracolactam C (**4.3**) was not detected in either dataset, likely due to compound degradation. Therefore, micromonolactam (**4.1**) and dracolactam A (**4.2**) were used as the target molecules.

The SYNAPT G2-Si data were consistent with earlier findings outlined in **Section 2.2.4**, with the GT target molecules identified in all mixtures. Analysis of the Xevo MRT preprocessed data showed that the masses of compounds **4.1** and **4.2** were detected in all mixtures. For compound **4.1**, the *m/z* 452.27905 was detected, corresponding to the water loss ($[M\text{-}H_2O+H]^+$, calc. *m/z* 452.2795, Δ -0.99 ppm). Progenesis QI calculated the exact mass of the molecule as 469.28239 (calc. 469.2828), with a ppm error of Δ -0.90. For compound **4.2**, the protonated ion cluster was detected at $[M+H]^+$ *m/z* 486.28451 (calc. *m/z* 486.2850), with a Δ of -1.00 ppm error. Compared to the SYNAPT G2-Si data, GT molecules **4.1** and **4.2** exhibited ppm errors of 1.12 and 2.30, respectively. Notably, these data maintained mass accuracy, detecting the molecule correctly in all mixtures up to *r* = 100.

When examining the microbial fractions added to the GT sample to increase background complexity, it was identified that fraction RLUS-2144D contained a known metabolite, surugamide A (**4.4**). The microbial fraction was added at *r* = 20, making surugamide A a suitable candidate to test instrument sensitivity for accurate annotation. In the *r* = 20 sample, surugamide A was detected at $[M+H]^+$ *m/z* 912.62845 (calc. *m/z* 912.6286, Δ 0.20 ppm) with excellent resolution on the Xevo MRT. By contrast, surugamide A (**4.4**) was detected on the SYNAPT G2-Si at $[M+H]^+$ *m/z* 912.6278 (calc. *m/z* 912.6286, Δ 0.92 ppm). The MS spectra differences between the two instruments are illustrated in **Figure 4.6**.

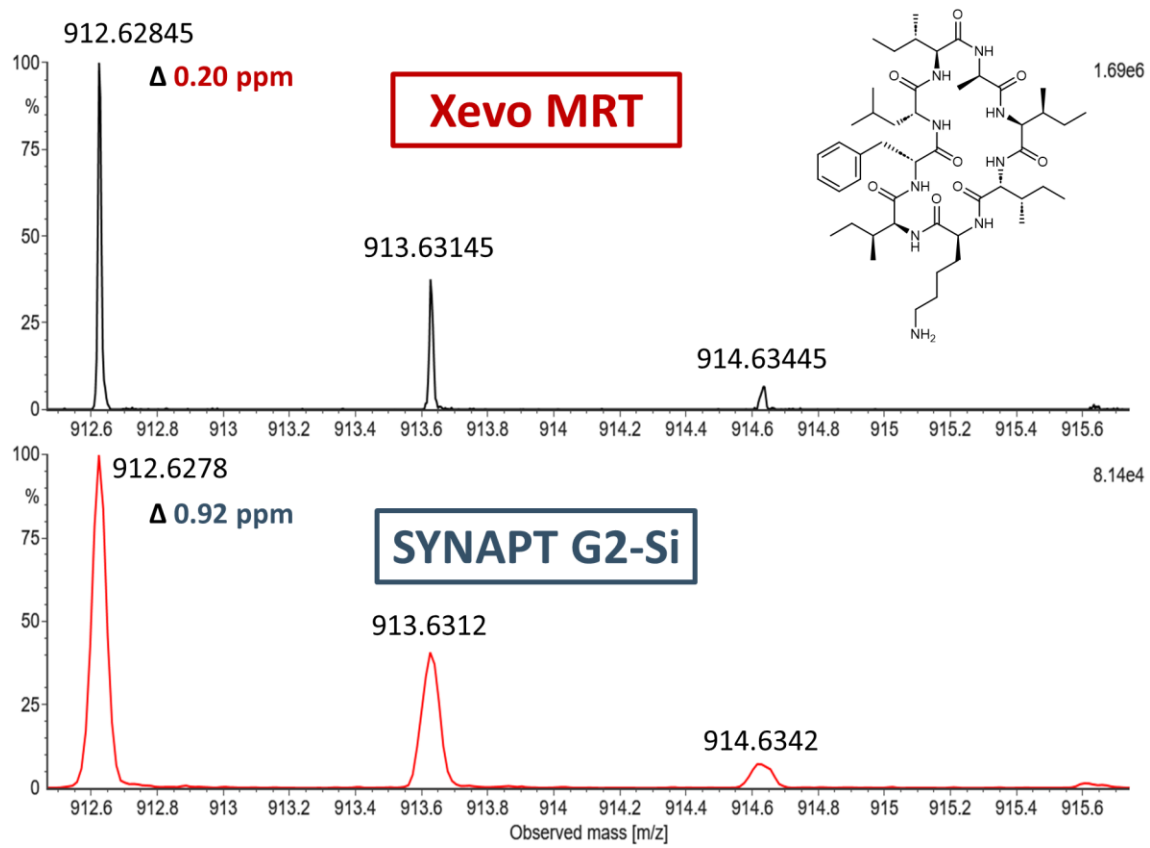


Figure 4.6 **Mass Spectrum Differences of Known Metabolite Surugamide A (4.4) Between the Xevo MRT and SYNAPT G2-Si.**
Surugamide A (**4.4**), a known NP isolated from a marine *Streptomyces* sp., is abundant in a fractionated extract (RLUS-2144D) from the Linington Lab library. In the mixture r = 20, **4.4** was detected with high mass accuracy (<1 ppm) on both test instruments. The top spectrum in the figure is from the Xevo MRT instrument, and the bottom spectrum is from the SYNAPT G2-Si. This figure highlights the difference in resolving power between the two instruments. It demonstrates the improved accuracy of measurements when using HR instruments to detect and annotate molecules in complex mixtures.

These results demonstrate that HR instruments with increased sensitivity detect relevant ions with high mass accuracy, even in increasingly complex sample matrices. The Xevo MRT outperformed the SYNAPT G2-Si, highlighting its superior measurement capabilities. This improvement allows researchers to annotate molecules with greater accuracy, significantly enhancing the discovery pipeline and proving crucial in identifying potentially thousands of molecules in a metabolomics dataset.

4.3.2. Performance of High-Sensitivity Mass Spectrometry in Complex Mixture Analysis.

The second part of this experiment was designed to assess instrument limitations when pooling many complex samples. Using more powerful instrumentation to analyze complex mixtures offers advantages such as increased sensitivity in detector technology, faster scanning speed, and improved resolving power. Researchers can leverage HR mass spectrometers to detect more MS features with higher mass accuracy when conducting pooled experiments. This contributes to higher-quality datasets and enhances the potential for pooling more samples, resulting in higher throughput. Today's advanced instrumentation, exemplified by the MMS platform, provides unparalleled data quality for efficiently analyzing large sample libraries and facilitating timely drug discovery.

The study employed two Waters instruments: the SYNAPT G2-Si and the new Xevo MRT. Both instruments feature highly sensitive detectors capable of identifying low-abundance MS features with high mass accuracy. The Xevo MRT exhibits a sensitivity increase of 3x compared to the Waters Xevo G3 qTOF instrument—one of the best-in-class qTOF mass spectrometers at its release. Examining MS feature retention across complex samples aimed to understand how well the instruments recall essential details about a sample during sample pooling experiments. These results also shed light on the limits of sample pooling within a given specification range.

The triplicate GT samples underwent replicate comparison, retaining only those MS features detected in at least 2 out of 3 samples. Once a GT dataset was established from each instrument analysis, the assessment focused on how well the GT MS features were recalled as a function of increased background complexity. After processing the data from the SYNAPT G2-Si MS, it was found that the GT dataset consisted of 283 MS features. In contrast, the Xevo MRT boasted 1064 GT features—a nearly four-fold increase in detected MS features. This increase is likely due to the increase in instrument sensitivity, which allows the detection of less abundant fragment ions, isotope peaks, and adducts in the sample set. Despite analyzing replicates of the same sample set, the Xevo MRT analysis occurred at the Waters Research and Development facility in Wilmslow, England. The same UPLC instrumentation was used in both experiments, and the sample preparation followed the same protocol in both cases. Still, variations

(such as retention time and MS feature lists) occurred despite closely matching sample preparation protocols. The two GT datasets were compared between instruments, with the SYNAPT G2-Si dataset serving as the benchmark. MS features were grouped based on recorded retention time and *m/z*, eliminating features before 1 minute (solvent front) and after 6 minutes (wash gradient) (**see Section 4.5.3**). Initial observations of the Xevo MRT dataset revealed a retention time difference of +0.27 minutes, which was corrected to align with the SYNAPT G2-Si. Of the 283 MS features in the SYNAPT G2-Si dataset, 186 of these features were identified in the Xevo MRT dataset despite detecting over 1000 MS features in the Xevo MRT experiment. For the GT molecules **4.1** and **4.2**, the number of MS features detected for each molecule was 19 and 15, respectively, according to Progenesis QI. In the Xevo MRT data, the GT molecules contained 26 and 22 MS features each, highlighting the additional sample information available when using more sensitive instrumentation.

As the experiment progressed and more complex mixtures were added to the GT sample, the analysis focused on how well the GT features were recalled from mixtures ranging from $r = 2$ to $r = 100$. For the SYNAPT G2-Si dataset, approximately 95% of the GT features could be recalled when pooling up to 10 samples, maintaining 83% of GT features up to 20 samples, and approximately 78% when pooling upwards of 30 samples. In the Xevo MRT dataset, overlapping MS features were assessed, revealing a significant improvement in MS feature recall. Even when pooling 50 samples, the Xevo MRT maintained an 85% GT MS feature recall (**Figure 4.7**).

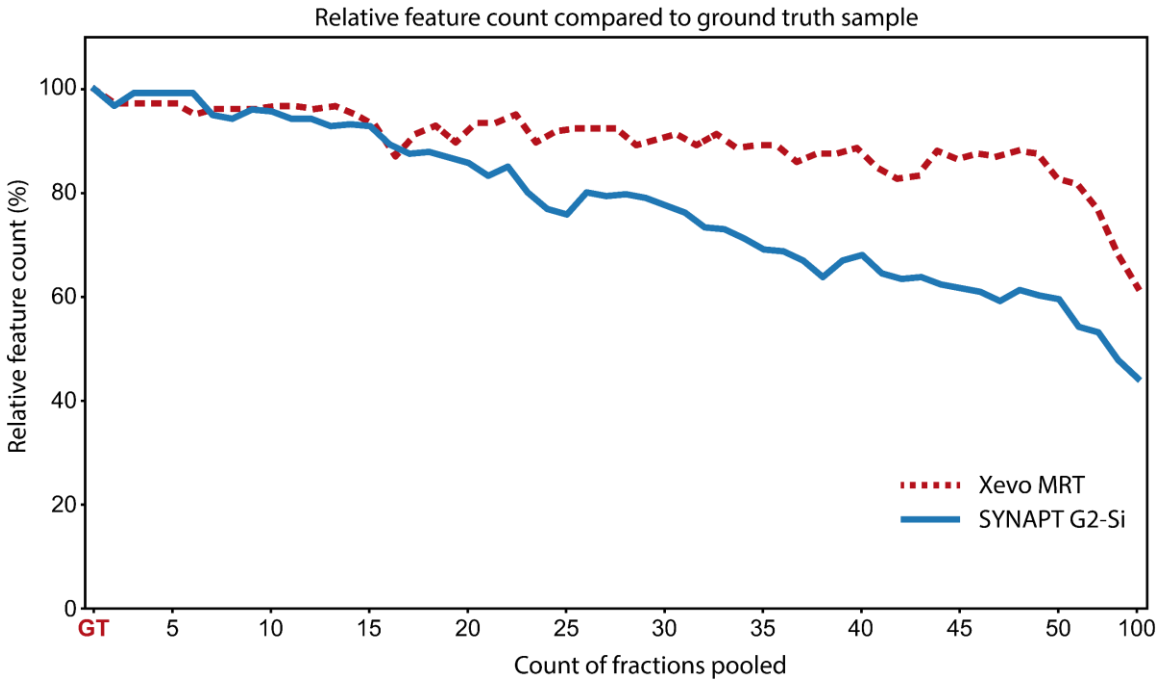


Figure 4.7 **Relative Feature Recovery of the Full Ground-Truth Feature Lists.** The GT feature lists were compared to assess the recall effectiveness of MS features from the GT dataset. The impact of increasing sample complexity was examined by pooling consecutively more samples together (from $r = 2$ to $r = 100$). The blue line represents MS feature recall from data obtained using the SYNAPT G2-Si. In contrast, the dotted red line represents the recall of overlapping MS features between datasets from the Xevo MRT. The SYNAPT G2-Si detects approximately 78% of GT MS features when pooling 30 samples. In contrast, the Xevo MRT – with heightened instrument sensitivity – improves data recall by detecting 85% of overlapping MS features, even when pooling up to 50 complex mixtures.

A thorough analysis of the loss in MS feature detection revealed that it was due to fragment ions and isotopologues in low abundance in the mixtures (**Figure 4.8**).

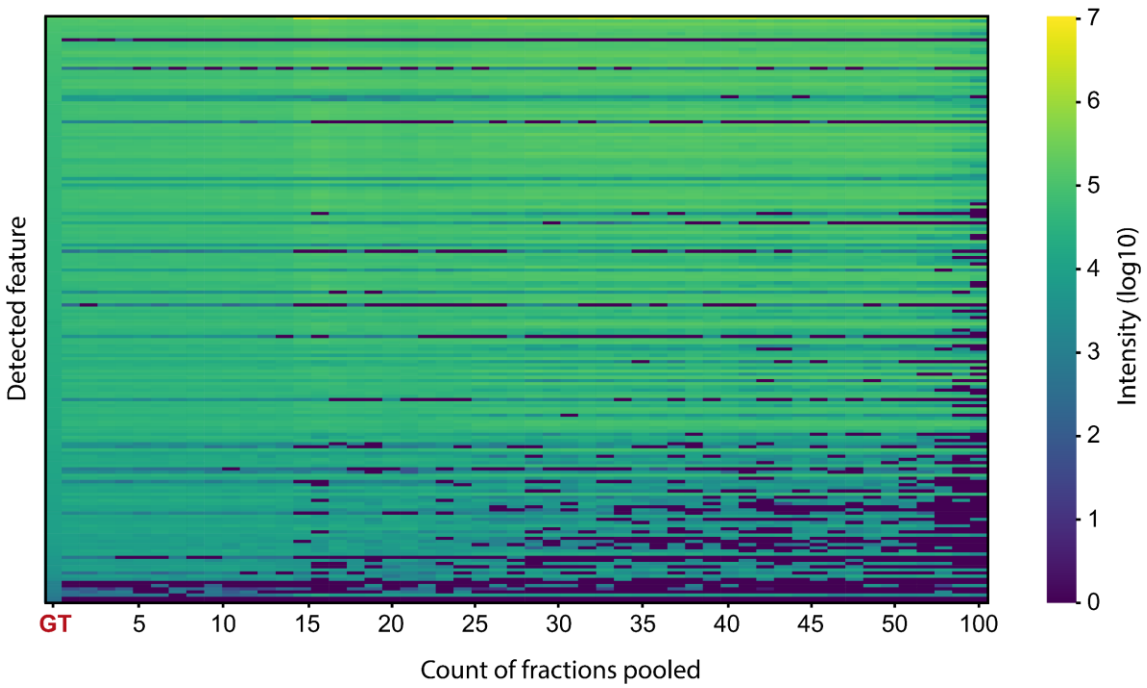


Figure 4.8 Assessment of MS Feature Recovery as a Function of Mixture Complexity of Data from the Xevo MRT.
Plotting GT MS features along the y-axis in order of feature intensity within the GT dataset, a thorough examination reveals which features remain undetected as the background complexity of the pooled samples increases. The analysis indicates that most of the MS features not recalled in the experiment were low-abundance fragment ions and isotopologues—data not crucial for bioactive molecule prioritization in downstream analyses.

This experiment highlights the significant improvement in MS feature detection resulting from the increased resolving power and sensitivity of the Xevo MRT. The objective was to assess how effectively the instrument could detect all the GT MS features. Investing in cutting-edge technology can significantly enhance throughput potential when analyzing large extract libraries using platforms like MMS. The number of samples pooled directly correlates with the fold-change improvement in analysis throughput and time savings during UPLC-MS analysis, all while maintaining high mass accuracy for detected analytes.

4.4. Conclusions

This study aimed to evaluate the impact of instrument choice on the quality of MS data in HT metabolomics, particularly as sample complexity increases. The research compared two HR-MS instruments, the Xevo MRT and the SYNAPT G2-Si, using a GT microbial fraction containing known metabolites. The addition of complex mixtures to the

GT sample tested the ability of each instrument to detect these molecules and maintain mass accuracy throughout the experiment. The results demonstrated that the Xevo MRT outperformed the SYNAPT G2-Si in detecting the GT MS features with higher mass accuracy. Specifically, the Xevo MRT identified nearly 4x more GT MS features than the SYNAPT G2-Si, showcasing superior instrument sensitivity. This increase in sensitivity allowed the Xevo MRT to detect lower abundance fragment ions, enhancing the overall quality and depth of the metabolomics data.

For specific molecules like micromonolactam (**4.1**) and dracolactam A (**4.2**), the Xevo MRT and SYNAPT G2-Si consistently detected these compounds with high mass accuracy (<5 ppm), even in highly complex mixtures. However, the Xevo MRT detected the compounds below one ppm mass error, compared to the SYNAPT G2-Si that detected the molecules with a 1.12 and 2.30 ppm mass error, respectively. The high mass accuracy of known NP surugamide A (**4.4**) at low concentrations further highlighted the Xevo MRT's advanced capabilities, with a mass error of 0.20 ppm (SYNAPT G2-Si = 0.92). The superior performance of the Xevo MRT is attributed to its enhanced sensitivity, which is reportedly 3x more significant than the already advanced Waters Xevo G3 qTOF. The pooling experiment highlighted the advantages of using HR-MS for analyzing complex samples. The Xevo MRT maintained a high recall rate of ground-truth (GT) features even as the sample complexity increased, retaining 85% of GT MS features when pooling up to 50 samples. In contrast, the SYNAPT G2-Si recall rate dropped significantly as the sample complexity increased, especially beyond pooling 20 samples.

This study demonstrates that investing in advanced MS technology, such as the Xevo MRT, can significantly enhance the speed and quality of metabolomics data analysis. The technology's improved sensitivity and resolving power help accurately identify molecules, speed up the discovery process and improve the efficiency of analyzing large sample libraries. These advancements are essential for accurately identifying molecules in complex metabolomics datasets, highlighting the importance of selecting the right instrument for high-throughput metabolomics research.

4.5. Materials and Methods

4.5.1. General Experimental Information

All solvents used in the MS acquisition were of optima LC-MS grade. Acetonitrile (ACN), methanol (MeOH), and formic acid (HCOOH) were purchased from Thermo Fisher Scientific. Deionized water was obtained using an 18 MΩ·cm Milli-Q system (EMD Millipore Corporation). HR-MS acquisition and metabolomics experiments were performed on a Waters Xevo MRT and SYNAPT G2-Si qTOF mass spectrometer.

4.5.2. UPLC-MS Acquisition Conditions

MS measurements were performed on a SYNAPT G2-Si (Vancouver, BC, Canada) and an Xevo MRT (Wilmslow, England), following the parameters outlined in **Section 2.4.2**.

4.5.3. Mass Spectrometry Data Processing

All samples were processed using the Progenesis QI software suite (v3.0.9293.38961). Lock-mass corrected positive data was uploaded into Progenesis QI for spectra alignment and peak picking using default settings. Positive polarity adducts were detected based on $[M+H]^+$, $[M+Na]^+$, $[M-H_2O+H]^+$, and $[M+K]^+$ ions. The datasets from the two instruments were processed separately because acquisition parameters differed, so Progenesis QI processing could not be performed together. Once both datasets were processed, the generated feature table, including the *m/z*, predicted neutral mass, retention time, and normalized intensity values, were exported for feature comparison between datasets. Strict blank subtraction was implemented in both datasets, whereby MS features in the solvent blank samples were omitted from the resulting MS feature table. The GT triplicates were replicated and compared, retaining 2 out of 3 detected features to establish the GT dataset. MS features that eluted before 1 minute and after 6 minutes were removed from the tables as an additional filter. Next, the feature table from the Xevo MRT exhibited a shift in retention time values by +0.27 minutes, based on the detection of the GT molecules **4.1** and **4.2**. Following the retention time correction, MS features from the SYNAPT G2-Si feature table were compared to the Xevo MRT dataset, detecting the GT features in a dataset of many

more MS features. GT features were grouped with Xevo MRT features within a two ppm mass error range and 0.1 minute retention time window.

Chapter 5.

Unraveling Complex Natural Products Structures Through Advanced Spectrometric and Chemical Techniques

5.1. Biological Relevance of the Burkholderiales Order

Developments in genome sequencing technologies have revolutionized our understanding of the genetic composition of various organisms. Genome sequencing has had significant implications for exploring the potential of bioactive natural products (NPs) that organisms may produce.[144] By analyzing genetic information, researchers can uncover pathways involved in synthesizing these molecules, which may have therapeutic applications.[145] A notable outcome of these technological advancements is the ability to investigate bacterial taxa beyond well-studied organisms. These bacteria may harbor genetic pathways responsible for synthesizing bioactive compounds with diverse properties.[144] *Burkholderia* is one example of a group of Gram-negative bacteria with diverse ecological niches. Despite its ambiguous taxonomic classification, the genus *Burkholderia* has garnered attention due to reports of its antibacterial, antifungal, and cytotoxic activities. The characteristic genetic makeup of *Burkholderia* species provides a rich source for discovering novel bioactive compounds.[146] Exploring the genetic pathways involved in NP production in *Burkholderia* allows researchers to identify potentially therapeutic molecules to combat various diseases.

Indeed, the ubiquity of *Burkholderia* across a broad spectrum of terrestrial and aquatic environments exposes it to a myriad of interactions, ranging from beneficial to antagonistic.[147] These dynamic ecological relations prompt *Burkholderia* to adapt and develop molecules to cope with such diverse pressures. The vast array of *Burkholderia*-produced NPs with varying environmental activities showcases its multifaceted interactions within different environments.[146] For example, *Burkholderia*'s ability to halt the cell cycle in rice cells, arising from its mutualistic relationship with pathogenic fungi, demonstrates its role in facilitating intricate interactions between plants and their microbial partners.[148–150] Similarly, *Burkholderia*-produced NPs with pesticidal activities against arthropod pests demonstrate their potential in integrated pest management

strategies.[151] These compounds also play a crucial role in various biological processes, including swarming, biofilm formation, iron acquisition, and quorum sensing.[152–155]

The biotechnological potential of *Burkholderia* extends beyond the production of NPs with environmental applications, encompassing a wide range of bioactive molecules with diverse molecular scaffolds and biological activities. One notable example of a successful NP derived from *Burkholderia* is FK228 (romidepsin), a United States Food and Drug Administration (FDA)-approved anticancer drug (**Figure 5.1**).[151] FK228 is a hybrid peptide with polyketide (PK) – nonribosomal peptide synthetase (NRPS) components, highlighting the intricate biosynthetic machinery employed by the bacterium to produce complex molecules with therapeutic potential. Other bioactive molecules from *Burkholderia* include spliceostatin A,[156] occidiofungin A – D,[157] and ornibactin,[158,159] each with distinctive pharmacological properties ranging from antifungal and antibacterial to anticancer activities (**Figure 5.1**).

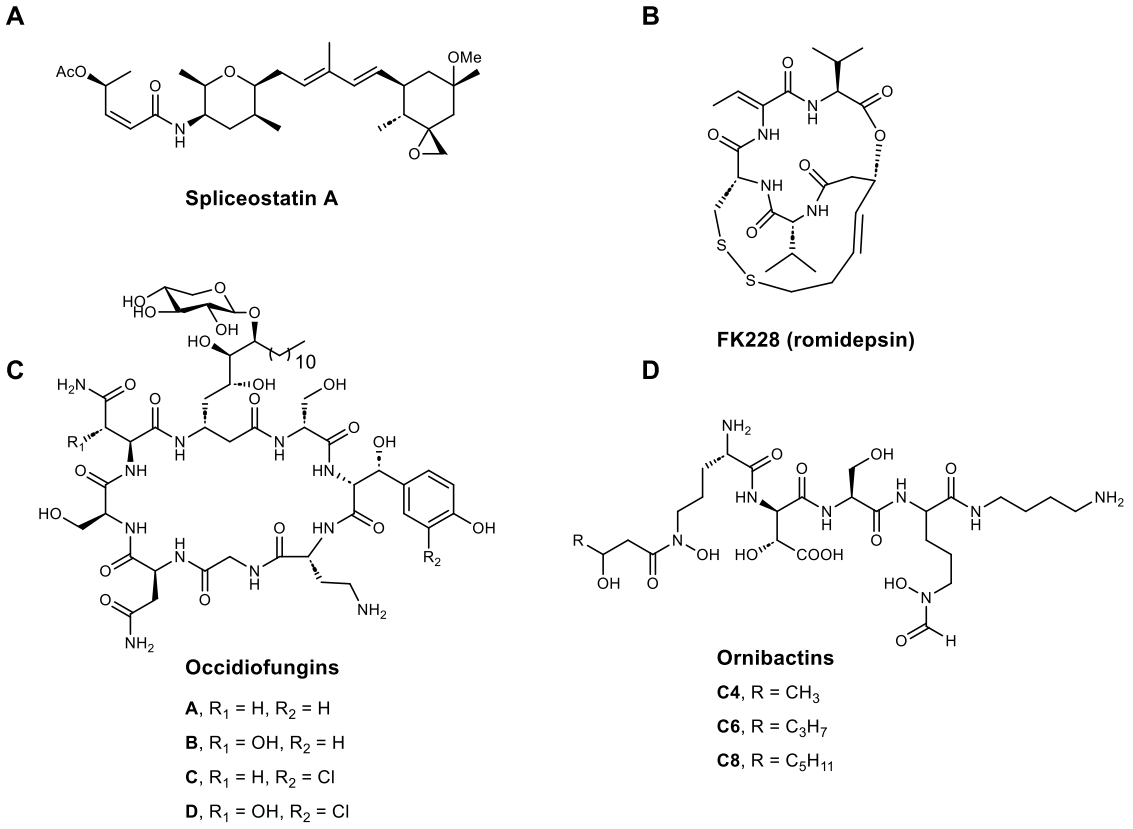


Figure 5.1 **Natural Products Isolated from *Burkholderia* sp.**
(**A**) Spliceostatin A, a semisynthetic analog belonging to a class of spliceosome inhibitors. These inhibitors are synthesized by a hybrid PK-NRPS system and evaluated in cell-proliferation assays against a panel of solid tumor cancer cell

lines;[156,160] (**B**) FK228 (romidepsin), an FDA-approved depsipeptide that acts as a natural histone deacetylase (HDACs) inhibitor to treat cutaneous and peripheral T-cell lymphoma;[151] (**C**) Occidiofungin A – D, glycopeptides with antifungal activity inhibiting a large spectrum of fungal pathogens;[157] (**D**) Ornibactins, NRPS-synthesized tetrapeptide siderophores with reported antibacterial activity.[158,159]

Genomic analysis of *Burkholderia* species has revealed the presence of biosynthetic gene clusters (BGCs) responsible for producing these NPs.[146] These genetic blueprints provide researchers with valuable insights into the molecular machinery involved in NP biosynthesis. Briefly, BGCs are comprised of specific regions that encode for various protein domains that are responsible for the assembly of NPs.[161] Deciphering these pathways permits researchers to predict the origin of specific NPs and explore the potential of silent or under expressed BGCs to discover novel compounds.[162] This approach, known as genome mining, has revolutionized the field of NP discovery by enabling the systematic exploration of microbial genomes for biosynthetic potential.

Silent or under expressed BGCs often hindered the discovery of novel molecules using traditional bioassay-guided discovery methods. This limitation arose because the molecules produced by these clusters would fall below the detection limit in target bioassays.[146] Genome mining has addressed these limitations by allowing researchers to systematically survey microbial genomes for BGCs of interest, irrespective of expression levels.[163,164] However, obtaining NPs from predicted BGCs remains a bottleneck because many BGCs are under expressed to produce substantial quantities of the NP to allow detection and isolation.[162] Advanced genetic engineering techniques provide avenues for accessing these molecules. Techniques such as biosynthetic investigations and promoter exchange can activate gene expression if the gene cluster is not transcribed in the native producer. Heterologous expression is another genetic engineering technique that has proven to be a valuable approach for NP discovery from *Burkholderia*, allowing to produce bioactive compounds in genetically tractable hosts.[165] Additionally, integrating genome sequencing technologies with bioinformatics tools has enhanced the confidence in predicting molecules from BGCs. This enables researchers to identify key features within BGCs, such as gene organization, sequence motifs, and enzymatic domains, providing valuable insights into the biosynthetic pathways and potential products encoded by these clusters.[166,167]

This chapter describes two studies investigating *Burkholderia* BGCs through a combination of genetic engineering and bioinformatic techniques to unlock the vast chemical diversity potential encoded within the microbial genome, leading to the discovery of novel NPs. **Section 5.3** focuses on the identification, isolation, and structure elucidation of a large lipodepsipeptide encoded within a giant NRPS.[168] **Section 5.4** explores the advantages of heterologous expression and promotor activation for producing NPs that were previously under expressed. These NPs were identified through genomic mining of a *Burkholderia* bacterium.

5.2. Identification of the Lipodepsipeptide Selethramide Encoded in a Giant Nonribosomal Peptide Synthetase from a *Burkholderia* Bacterium

Published manuscript

Reprinted with permission from *Proc. Natl. Acad. Sci. U. S. A.* **2023** *120*(42), e2304668120. DOI: 10.1073/pnas.2304668120.

Authors: Sean B. Romanowski, Sanghoon Lee, Sylvia Kunakom, Bruno S. Paulo S. Paulo, Michael J. J. Recchia, Dennis Y. Liu, Hannah Cavanagh, Roger G. Linington, Alessandra S. Eustáquio

Author Contributions: R.G.L. and A.S.E. designed the study. S.B.R., S.L., S.K., B.S.P., M.J.J.R., D.Y.L., and H.C. performed research. S.B.R., S.L., S.K., B.S.P., M.J.J.R., D.Y.L., H.C., R.G.L., and A.S.E. analyzed data. A.S.E. authored the paper with input from all authors. Specifically, M.J.J.R. contributed to the structure elucidation critiques from reviews following submission. M.J.J.R. reanalyzed amino acid configurations using Marfey's analysis with a more robust chromatographic method to separate amino acid derivatives. Secondly, reviewers questioned the lactone ring location. M.J.J.R. reevaluated the NMR data to confirm lactone linkage. To confirm the amino acid sequence within the molecule, M.J.J.R. performed base hydrolysis on the molecule to open the lactone and performed MS/MS experiments to confirm the sequence. Finally,

despite attempting to determine positional configurations within the molecule due to the repeating sequence of leucine and serine molecules, this effort was unsuccessful. Instead, the information was derived from an in-depth domain analysis of the BGC by B.S.P.

5.2.1. Introduction

This section discusses the isolation and structure elucidation of a genetically encoded NP with therapeutic potential in medicine and agriculture. As mentioned earlier, the emergence of advanced bioinformatics tools for genome mining has markedly enhanced the prediction of BGCs responsible for encoding novel NPs.[167] *Burkholderia* bacteria have emerged as a prolific source of NPs, exhibiting substantial chemodiversity and demonstrating activity across diverse biological targets.[146] However, despite the abundant metabolic profile of the Burkholderiales order, it only represents a small fraction of the potential predicted from genomic studies.[169]

Obtaining NPs from predicted BGCs remains a significant challenge in the discovery process. This is due to insufficient quantities of NP produced by under expressed BGCs, making isolation difficult.[162] However, the metabolic potential of *Burkholderia* is promising. Multiple NPs are often predicted to be encoded in a large genome[170–172], resulting from advantageous interactions.[171] Exploring these silent or under expressed BGCs can expand the known NP chemical space.

Burkholderia sp. FERM BP-3421, obtained from terrestrial soil, has been identified as a prolific producer of bioactive metabolites (**Figure 5.1**).[151,156–160] This strain synthesizes abundant bioactive NPs and possesses diverse biosynthetic machinery for which no products are known. These include autologous PK-NRPS, polyketide synthetases (PKS), and ribosomally synthesized and post-translationally modified peptide (RiPP) pathways—additionally, *Burkholderia* sp. FERM BP-3421 can serve as a host organism for heterologously expressed genomes, potentially leading to higher yields of metabolites.[146] As a result, this strain has been extensively studied to explore its biosynthetic potential using various multi-omic approaches.

This study used a suite of multi-omic approaches, including genomics, metabolomics, and epigenomics, to reveal previously identified bioactive NPs and led to

the identification of a novel octadecalipodepsipeptide named selethramide (**5.1**, **Figure 5.2**). This new NP was encoded in a giant NRPS. Harnessed epigenomic data improved the deoxyribonucleic acid (DNA) transfer efficiency into FERM BP-3421, allowing the generation of a selethramide synthetase knockout mutant and identifying the molecule's role in swarming motility as a surfactant molecule. This model, which uses multi-omics and bioinformatics, allows for potential future studies of NPs. While this section will not outline the full description of the paper, it will highlight personal contributions to the project. This large molecule proved complex to assemble completely, requiring sophisticated spectroscopic and semisynthetic techniques to unambiguously assign the absolute configuration of amino acids, establish the ring system with the lactone linkage, and position the amino acids within the molecule.

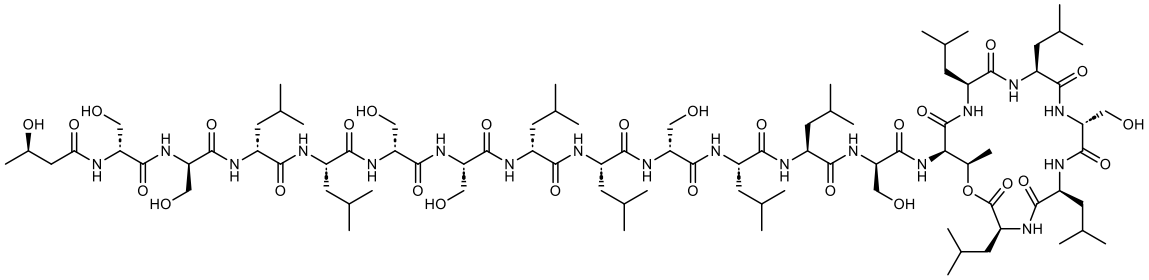


Figure 5.2 The Molecular Structure of Selethramide (5.1), an Octadecalipodepsipeptide Isolated From a *Burkholderia* sp. FERM BP-3421 Soil Isolate.

5.2.2. General Experimentation and Isolation of Selethramide (5.1)

It should be noted that S.L. performed the original isolation of **5.1** from *Burkholderia* sp. FERM BP-3421. Following the initial analysis, the target molecule required further isolation, necessitating additional bacterium culture for structural studies. Selethramide (**5.1**) was discovered based on the doubly charged ion cluster $[M+2H]^{2+}$ *m/z* 965 from the orphan metabolite features. The fragmentation pattern suggested that **5.1** was a large lipopeptide, indicated by a series of neutral losses corresponding to the molecular weights of serine and leucine/isoleucine residues. The fragmentation pattern culminated in a sizeable cyclic structure with an $[M+H]^+$ *m/z* 641. High-resolution mass spectrometry (HR-MS) predicted the molecular formula as $C_{89}H_{158}N_{18}O_{28}$ based on the protonated ion $[M+H]^+$ *m/z* 1928.1532. Genomic analysis of the *Burkholderia* sp. FERM BP-3421 strain revealed that the only BGC plausibly associated with **5.1** was the NRPS BGC 1.2. According to predictions, **5.1** consisted of an acyl chain with 18 amino acid

residues: six units of D-Ser, one unit of L-Ser, one unit of D-Thr/D-allo-Thr, eight units of L-Leu, and two units of D-Leu, which matched well with the domain architecture from the BGC.

Selethramide (**5.1**) was isolated as an amorphous white solid from a liquid culture derived from a frozen bacterial stock of *Burkholderia* sp. FERM BP-3421. The following is a summary of the isolation process. First, the bacterium was inoculated onto lysogeny broth (LB) agar and incubated at 30 °C. The culture was scaled up to a larger volume (2 L) in a 2S4G medium, which provided sufficient material for subsequent extraction. The bacteria were then extracted using an organic solvent system consisting of dichloromethane (DCM) and methanol (MeOH). After extraction, the material was vacuum filtered to separate the solvent-extracted material and concentrated *in vacuo*. Next, the dried extract was subjected to primary fractionation using an aqueous methanol mobile phase and a C_{18} stationary phase in a stepwise gradient system, increasing by 20% aq. MeOH increments to produce seven fractions (A → F).

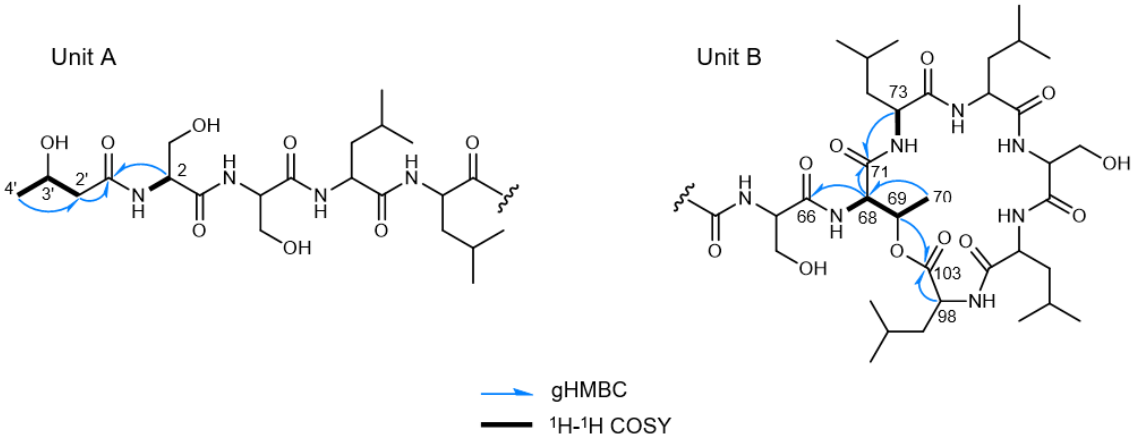
MS analysis of the primary fractions revealed that **5.1** was abundant in fraction 'E' (100% MeOH), which was subjected to reversed-phase (RP) high-pressure liquid chromatography (HPLC) to afford two subfractions E-1 and E-2 (VyDAC, 250 mm x 4.6 mm, 5 µm, 1.0 mL/min, 10% aq. MeOH to 95% aq. MeOH + 0.02% formic acid). E-1 was selected for further separation using analytical RP-HPLC in an isocratic elution system (Phenomenex Kinetex XB-C_{18}, 250 x 4.6 mm, 5 µm, 1.0 mL/min, 53% aq. ACN + 0.02% formic acid) to produce 5.81 mg of purified selethramide (**5.1**).

5.2.3. Contribution to Final Structure Determination

S.L. completed the original structure elucidation of selethramide (**5.1**). However, the initial reported structure faced questions during the review process. Additional experiments were conducted to refine and clarify the molecular structure and address the reviewed critiques.

Based on the predicted structure of **5.1** from BGC 1.2 (as shown earlier), performing MS/MS analysis on the native molecule would reveal its amino acid sequence. The MS/MS fragmentation data indicated that a portion of the molecule was cyclized, making it challenging to complete the amino acid sequence due to the

unpredictable fragmentation of cyclized peptides. Additionally, severe overlap in NMR signals precluded complete 2D structure elucidation of the molecule. The cyclization position was deduced from ^{1}H-^{1}H COSY cross-peaks between H-68 and H-69, as well as H-69 and H-70. HMBC correlations from H-69 and H-98 to C-103 provided further evidence of a cyclized structure extending from ^{13}Thr-69 to the terminal amino acid, ^{18}Leu-103, via one of the free hydroxy groups. This cyclization position was further confirmed by examining the structure of the basic hydrolysis product (details below).



Position	Unit A		Position	Unit B	
	δ_H (ppm)	δ_C (ppm)		δ_H (ppm)	δ_C (ppm)
2	4.47	53.3	66	--	175.2
1'	--	174.6	68	4.35	59.2
2'	2.39 2.47	45.9	69	5.24	72.3
			70	1.37	18.6
3'	4.20	66.0	71	--	171.3
4'	1.24	23.2	73	4.17	54.9
			98	4.00	62.8
			103	--	173.6

Figure 5.3 Selected 2D NMR Correlations and Chemical Shifts For Selethramide.

To comprehensively analyze the MS/MS fragmentation of the peptide, the lactone ring would need to be liberated using base hydrolysis. This process breaks the lactone linkage and adds a methoxy group (+32 Da) to the carboxyl end of the terminal amino acid, linearizing the molecule and allowing MS/MS experimentation to reveal the

complete sequence of amino acids. To break down the cyclic portion of the peptide, a sample of selethramide (0.1 mg) was suspended in MeOH (0.5 mL), and then a solution of NaOMe (20 µL, 0.5M in MeOH) was added and stirred at room temperature. MS was used to monitor the conversion until all the initial material had transformed into the base hydrolysis product. Subsequent MS/MS experiments were conducted on the base hydrolysis product using a SYNAPT G2-Si UPLC-ESI-qTOF mass spectrometer (Waters) analyzed in data-dependent acquisition mode. The instrument was equipped with a Waters ACQUITY *i*-Class UPLC and an HSS T3 column (2.1 × 100 mm, 1.8 µm). MS analysis of the derivative selethramide methyl ester product showed a $[M+H]^+$ *m/z* 1960.1804 (calc'd 1960.1833), and an absence of the starting material shows complete conversion to the product. The amino acid sequences with major fragments consistent with the losses of the acyl chain-Ser1 (y_{17} ion, calc'd 1787.1134, expt 1787.1047), Ser2 (y_{16} ion, calc'd 1700.0819, expt 1700.0661), Leu3 (y_{15} ion, calc'd 1586.9979, expt 1586.9896), Leu4 (y_{14} ion, calc'd 1473.9138, expt 1473.9160), Ser5 (y_{13} ion, calc'd 1386.8818, expt 1386.8784), Ser6 (y_{12} ion, calc'd 1299.8488, expt 1299.8474), Leu7 (y_{11} ion, calc'd 1186.7657, expt 1186.7670), Leu8 (y_{10} ion, calc'd 1073.6816, expt 1073.6794), Ser9 (y_9 ion, calc'd 986.6496, expt 986.6479), Leu10 (y_8 ion, calc'd 873.5655, expt 873.5634), Leu11 (y_7 ion, calc'd 760.4815, expt 760.4791), Ser12 (y_6 ion, calc'd 673.4495, expt 673.4478), Thr13 (y_5 ion, calc'd 572.4018, expt 572.3998), Leu14 (y_4 ion, calc'd 459.3178, expt 459.3181), Leu15 (y_3 ion, calc'd 346.2336, expt 346.2321), Ser16 (y_2 ion, calc'd 259.2016, expt 259.2002), and Leu17 (y_1 ion, calc'd 146.1176, expt 146.1174) (**Figure 5.4**).

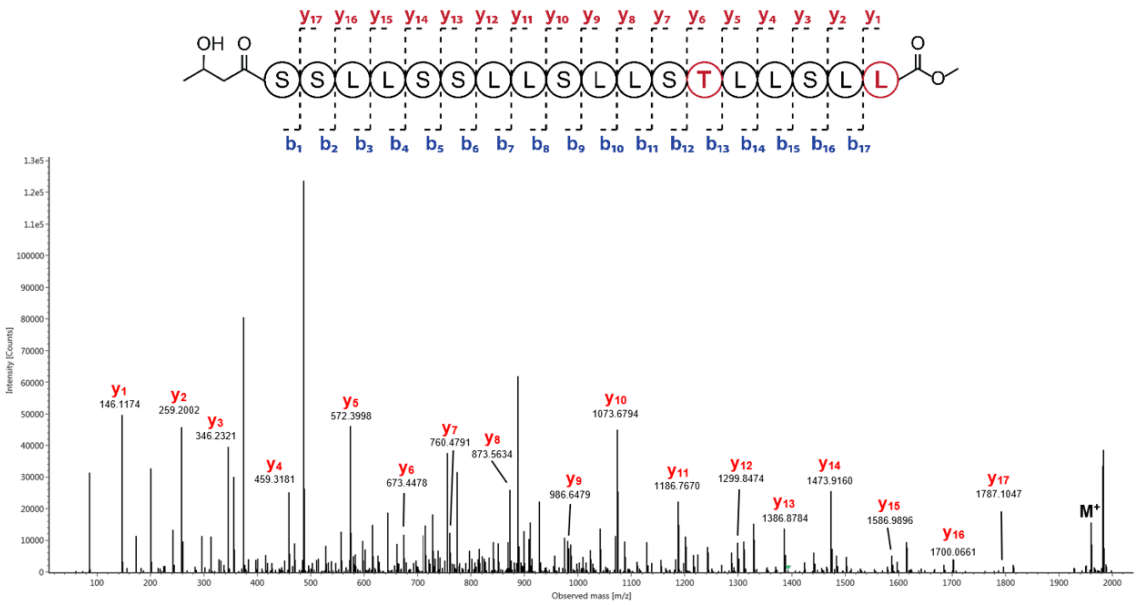


Figure 5.4 MS/MS Fragmentation Pattern of Base Hydrolysis Selethramide Product.
The amino acid sequences with major fragments consistent with the losses of the acyl chain-Ser1 (y_{17} ion, calc'd 1787.1134, expt 1787.1047), Ser2 (y_{16} ion, calc'd 1700.0819, expt 1700.0661), Leu3 (y_{15} ion, calc'd 1586.9979, expt 1586.9896), Leu4 (y_{14} ion, calc'd 1473.9138, expt 1473.9160), Ser5 (y_{13} ion, calc'd 1386.8818, expt 1386.8784), Ser6 (y_{12} ion, calc'd 1299.8488, expt 1299.8474), Leu7 (y_{11} ion, calc'd 1186.7657, expt 1186.7670), Leu8 (y_{10} ion, calc'd 1073.6816, expt 1073.6794), Ser9 (y_9 ion, calc'd 986.6496, expt 986.6479), Leu10 (y_8 ion, calc'd 873.5655, expt 873.5634), Leu11 (y_7 ion, calc'd 760.4815, expt 760.4791), Ser12 (y_6 ion, calc'd 673.4495, expt 673.4478), Thr13 (y_5 ion, calc'd 572.4018, expt 572.3998), Leu14 (y_4 ion, calc'd 459.3178, expt 459.3181), Leu15 (y_3 ion, calc'd 346.2336, expt 346.2321), Ser16 (y_2 ion, calc'd 259.2016, expt 259.2002), and Leu17 (y_1 ion, calc'd 146.1176, expt 146.1174).

After establishing the amino acid sequence in the linear peptide derivative, the focus shifted to determining the absolute configuration of the side chains. This determination is achieved through Marfey's analysis, which involves reacting liberated amino acids from the peptide (obtained via acid hydrolysis) with Marfey's reagent to form diastereomers. This process allows the separation of amino acid isomers by RP chromatography, providing a facile way to assess absolute configuration. The derivatized amino acids are then compared to both L and D laboratory amino acid standards. By determining the retention times of these standards, we can deduce the configuration of the amino acids within the molecule.

An aliquot of selethramide (0.3 mg) was hydrolyzed with 6N HCl at 110 °C for 18 hours, then dried under nitrogen gas. The resulting hydrolysate was resuspended in

50% aq. acetone and derivatized with Marfey's reagent Nα-(2,4-dinitro-5-fluorophenyl)-L-valinamide (L-FDVA). The reaction was quenched and analyzed by UPLC-MS (Waters RDa TOF) equipped with an ACQUITY UPLC and HSS T3 column (2.1 x 100 mm, 1.8 µm). The gradient solvent system eluted from 5% aq. ACN to 95% aq. ACN + 0.01% formic acid over 13 minutes at a 0.5 mL/min flow. A library of 36 amino acid standards (L and D pairs) was prepared and analyzed like selethramide and saved as a digital library. The derivatives from selethramide hydrolysate were matched with the amino acid standards library, and the configuration of the side chains in the structure were assigned using Waters UNIFI software. Marfey's analysis revealed that **5.1** contained a combination of L and D serines (**Figure 5.5**), a D-*allo*-threonine (**Figure 5.6**), and L and D leucines (**Figure 5.7**). For a complete description of the procedure, **see Section 5.5.2**.

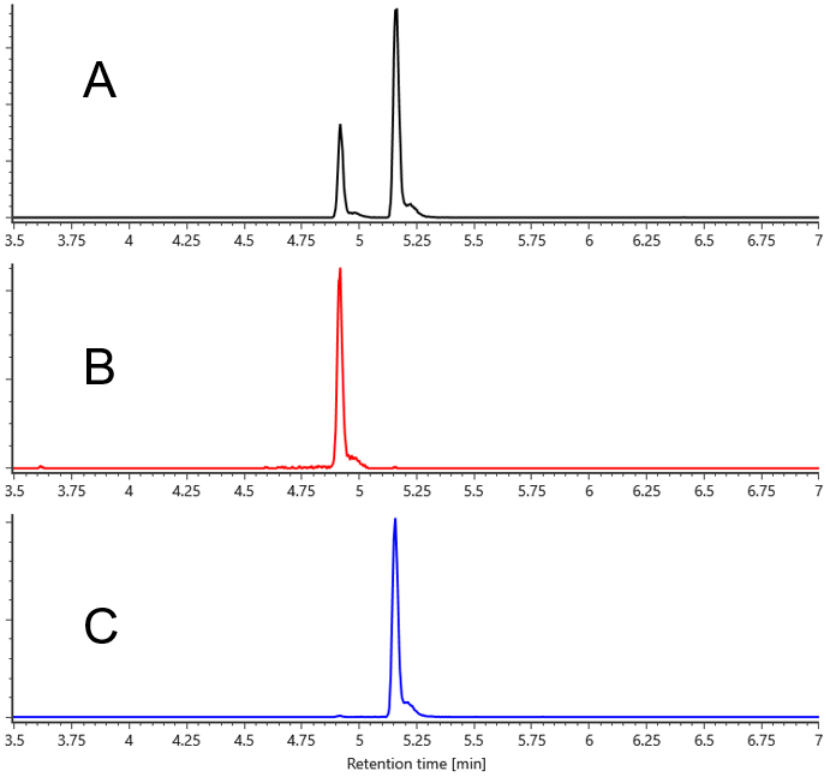


Figure 5.5 Extracted Ion Chromatogram for Marfey's FDVA-Derivatized Serine Hydrolysates.
(**A**) Selethramide derivatized hydrolysate EIC *m/z* 386.13. (**B**) FDVA-L-serine EIC *m/z* 386.13; RT = 4.92 min. (**C**) FDVA-D-serine EIC *m/z* 386.13; RT = 5.16 min.

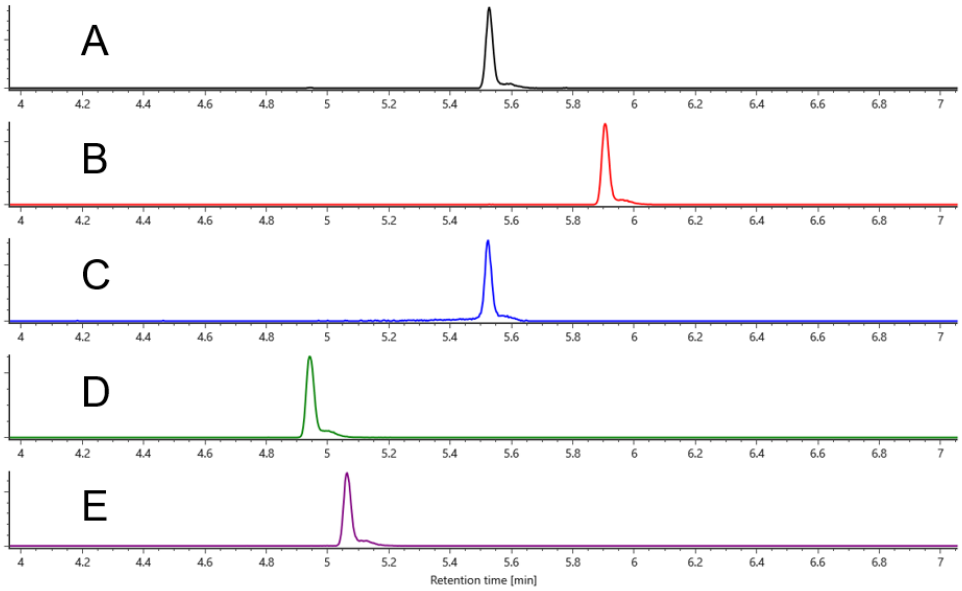


Figure 5.6 Extracted ion chromatograms for Marfey's FDVA-derivatized threonine hydrolysates.
(**A**) Selethramide derivatized hydrolysate EIC *m/z* 400.15. (**B**) FDVA-D-threonine EIC *m/z* 400.15; RT = 5.90 min. (**C**) FDVA-D-*allo*-threonine EIC *m/z* 400.15; RT = 5.52 min. (**D**) FDVA-L-threonine EIC *m/z* 400.15; RT = 4.94 min. (**E**) FDVA-L-*allo*-threonine EIC *m/z* 400.15; RT = 5.06 min.

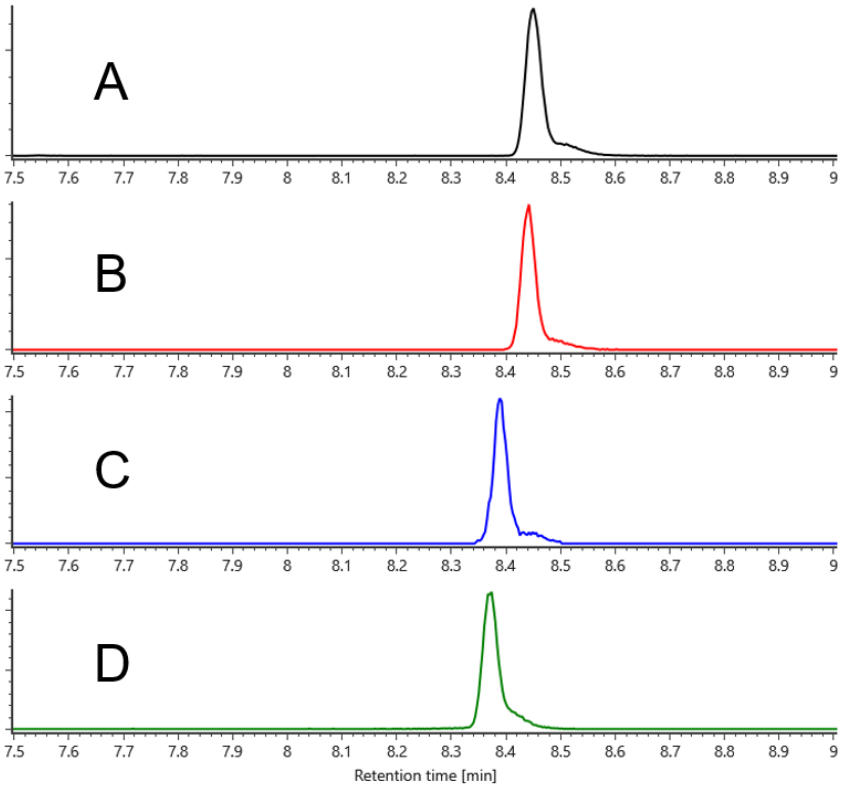


Figure 5.7 Extracted ion chromatograms for Marfey's FDVA-derivatized leucine hydrolysates.
(**A**) Selethramide derivatized hydrolysate EIC *m/z* 412.18. (**B**) FDVA-D-leucine EIC *m/z* 412.18; RT = 8.45 min. (**C**) FDVA-D-isoleucine EIC *m/z* 412.18; RT = 8.39 min. (**D**) FDVA-D-*allo*-isoleucine EIC *m/z* 412.18, RT = 8.37 min.

Since the sequence of amino acids in **5.**1 is known, the next task was to determine the configuration of each amino acid within that sequence. To achieve this, we conducted a feeding study in which the strain was exposed to deuterated serine and leucine. If an epimerization domain existed, we hypothesized that the incorporated amino acids would revert to their undeuterated forms. By analyzing the resulting MS/MS sequence, we could infer the configuration of each amino acid. Unfortunately, attempts to incorporate labeled amino acids through feeding studies did not yield the expected results. Attention was turned to identifying epimerization domains within the BGC through domain analysis. The analysis revealed that certain amino acids were in the D configuration. Combining this information with predictions from the BGC analysis, we assigned the absolute configuration of the amino acids as follows: D-Ser1-D-Ser2-D-Leu3-L-Leu4-D-Ser5-L-Ser6-D-Leu7-L-Leu8-D-Ser9-L-Leu10-L-Leu11-D-Ser12-D-*allo*-Thr13-L-Leu14-L-Leu15-D-Ser16-L-Leu17-L-Leu18 (**Figure 5.2**).

5.3. Discovery of Megapolipeptins by Genome Mining of a Burkholderiales Bacteria Collection

Manuscript in preparation

Authors: Bruno S. Paulo, Michael J. J. Recchia, Sanghoon Lee, Claire Fergusson, Sean B. Romanowski, Antonio Hernandez, Nyssa Krull, Dennis Y. Liu, Hannah Cavanagh, Allyson Bos, Christopher Gray, Brian T. Murphy, Roger G. Linington, Alessandra S. Eustáquio.

Author contributions: C.G. mentored A.B., B.T.M. mentored A.H. and N.K., R.G.L. mentored M.J.J.R., S.L., C.F., D.Y.L., and H.C., and A.S.E. mentored B.S.P. and S.B.R. R.G.L., and A.S.E. secured research funding. A.S.E. and R.G.L. designed the project, and C.F. isolated Burkholderiales strains from environmental samples. Subsequent strain analysis was performed by S.B.R., A.H., and N.K. using IDBac. B.S.P. isolated genomic DNA, conducted genome analyses, and identified candidate features. D.Y.L. and H.C. conducted antibacterial assays, and A.B. conducted the antifungal activity assays. M.J.J.R. and S.L. were responsible for isolating and structurally characterizing megapolipeptins. M.J.J.R.'s contributions included complete structure elucidation, Marfey's analysis to determine amino acid configurations, and NMR chemical shift analysis. Additionally, M.J.J.R. performed partial hydrolysis to ascertain the order of the threonine amino acid groups in the molecule. The collaborative effort extended to manuscript preparation, with B.S.P., M.J.J.R., R.G.L., and A.S.E. contributing to writing and reviewing the paper.

5.3.1. Introduction

Recent estimates suggest that only 3% of genome-predicted bacterial NPs have been successfully isolated and structurally characterized.[169] However, obtaining these NPs remains challenging because "silent" BGCs are not expressed in practical quantities for detection and isolation.[162] Despite this limitation, exploring the untapped biosynthetic potential of these underexplored NPs could have significant ecological implications in pharmaceutical and agricultural applications. This section explores the discovery of two novel NPs from predicted BGCs, achieved through activation approaches such as

engineering the native producer and heterologous expression in an optimized *Burkholderia* host strain.[173]

The heterologous expression of biosynthetic pathways for NPs is gaining popularity in microbial technology, drug discovery, and optimization.[174] This approach can streamline the discovery process, increasing product yields and improving throughput.[173] Beyond NP production, heterologous expression enables the creation of novel analogs through biosynthetic engineering. However, the choice of host strain can significantly influence the success of heterologous expression.[175] Recent studies have highlighted the connection between product yields and the genetic makeup of host strains and source DNA. This was demonstrated in the Eustáquio Lab, where researchers developed a *Burkholderia* host strain FERM BP-3421, which showed significantly higher production of native NPs than traditional *Escherichia coli* hosts.[176]

This study highlights a multifaceted approach to establishing a pipeline for discovering NPs from the Burkholderiales order. Leveraging a Burkholderiales culture collection gathered over five years from southern British Columbia, researchers employed IDBac – a matrix-assisted laser desorption/ionization (MALDI) TOF MS analysis method – to prioritize strains based on metabolic networks and maximize diversity.[177–179] Out of 230 strains analyzed by IDBac, 100 were selected for genome sequencing, with prioritization guided by biosynthetic classification tools such as antiSMASH (v6.0).[67] The study focused on a hybrid PKS-NRPS BGC conserved in the *Paraburkholderia megapolitana* clade, which includes polyunsaturated fatty acid (PUFA) genes. This decision was based on the fact that this type of BGC synthesizes molecules with the most potential for bioactivity from Burkholderiales.[146] This BGC was named *mgp* for megapolipeptin and was discovered in the in-house strain *P. megapolitana* RL18-039-BIC-B (genome #76).

Previous attempts to isolate these NPs from the strain *P. megapolitana* DSM 23488 were unsuccessful due to low yields, preventing isolation.[180] Interestingly, the BGC responsible for these NPs was silent in the wild-type strain but could be activated through promoter replacement. Despite detecting the *mgp* BGC known for producing compounds in the in-house strain RL18-039-BIC-B, no molecules were identified. Following previous promoter replacement strategies, the *mgp* BGC was cloned and heterologously expressed in the *Burkholderia* host strain FERM BP-3421.[168] Successful

incorporation into the host genome allowed us to identify two novel PKS-NRPS peptides, megapolipeptin A (**5.2**) and megapolipeptin B (**5.3**), containing varying fatty acid components (**Figure 5.8**). A thorough analysis using the NP Atlas showed low structural similarity to other known bacterial NPs (with maximum Tanimoto similarity scores of 0.58 (**5.2**) and 0.56 (**5.3**)).[53] This study highlights the advantages of a multi-omic approach and synthetic biology-enabled methods for discovering NPs from previously under expressed BGCs. While many researchers contributed to this project, this section will focus on personal contributions related to isolating and determining the structure of **5.2** and **5.3**.

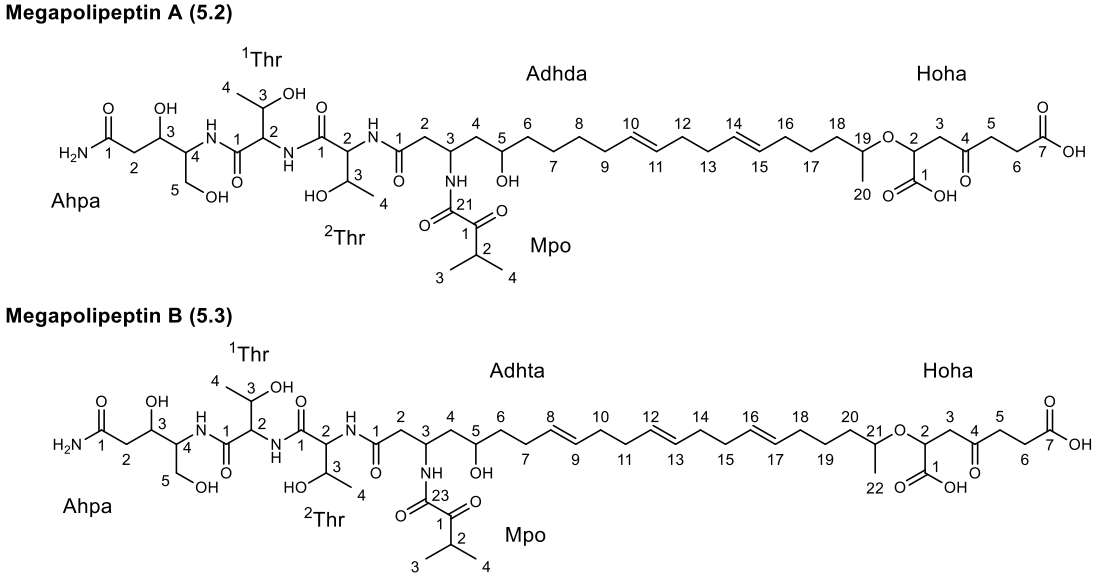


Figure 5.8 **Molecular Structures of NPs Isolated from P. megapolitana.**
The planar structures of megapolipeptin A (**5.2**) and megapolipeptin B (**5.3**).

5.3.2. Isolation and Structure Elucidation of Megapolipeptins A and B

The strain *Burkholderia* sp. FERM BP-3421 Δ*fr9A* (pBS2001) was genetically modified to express the *mgp* BGC. After fermentation, the strain was extracted using a mixture of DCM and MeOH in a 2:1 ratio. The resulting organic extract was concentrated and then subjected to primary fractionation using C_{18} solid-phase extraction (SPE) chromatography, with a stepwise elution gradient of MeOH/H_2O starting at 20% aq. MeOH to a 100% ethyl acetate wash for six fractionated samples (A → F). MS analysis of the fractions showed that fraction 'C' (60% aq. MeOH) contained the target molecules. Fraction 'C' was subjected to semi-preparative RP-HPLC (Phenomenex Synergy FUSION RP-80A, 250 x 10 mm) using an elution gradient of 10% aq. ACN to 80% aq.

ACN + 0.02% formic acid over 30 minutes to afford two fractions: C-1 (*m/z* 958, 19.3 minutes) and C-2 (*m/z* 984, 20.5 minutes). The two molecules were then purified using analytical RP-HPLC (Phenomenex Kinetex XB-C_{18}, 250 x 4.6 mm, 5 µm) using an isocratic elution system of 35% aq. ACN at 1.2 mL/min yielded 3.7 mg of compound **5.2** (t_R 5.20 minutes) and 9.8 mg of compound **5.3** (t_R 9.46 minutes). HR-MS data was acquired on a new Waters Multi-Reflecting Time-of-flight (MRT) mass spectrometer boasting a resolving power of 300,000 FWHM with parts-per-billion mass accuracy. Below is a compilation of the structure elucidation details for megapolipeptin A (**5.2**) and B (**5.3**) (**Figure 5.8**).

Megapolipeptin A (**5.2**) was obtained as a white powder. HR-MS data indicated a molecular formula of $C_{45}H_{75}N_5O_{17}$ based on the protonated cluster ion at $[M+H]^+$ *m/z* 958.52350 (calc'd. *m/z* 958.52307, 0.449 ppm) (**Appendix Figure A.17**). The tandem MS/MS spectrum revealed two sequential neutral losses consistent with threonine amino acid residues (**Figure 5.9**).

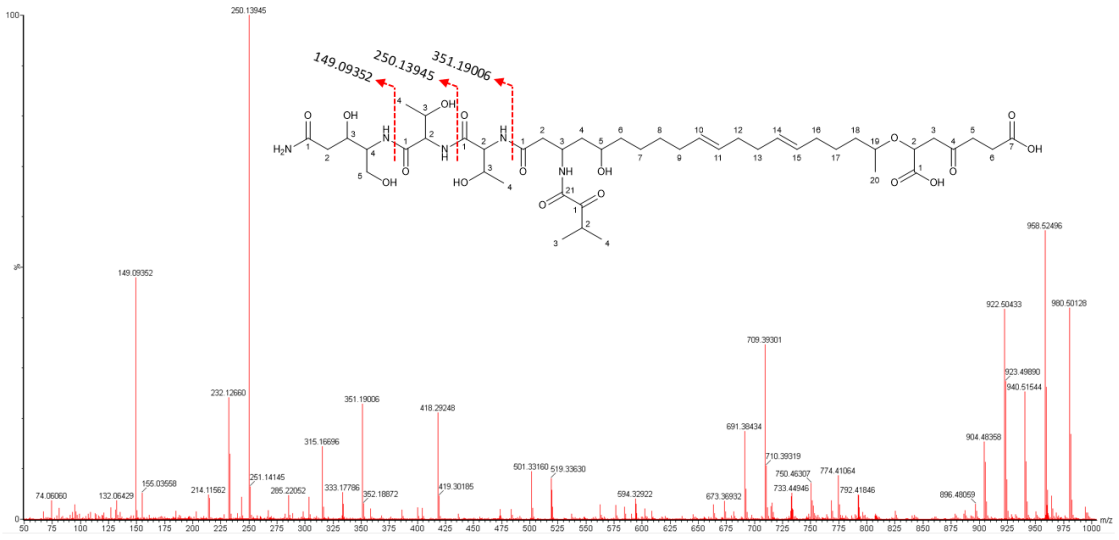


Figure 5.9 MS/MS Data for Megapolipeptin A (5.2).

To further analyze the structure, a comprehensive set of 1D (^{1}H and ^{13}C) and 2D (COSY, TOCSY, ROESY, NOESY, HSQC-DEPT, HMBC) NMR experiments were conducted to determine the planar structure recorded in DMSO-d_6 (**Appendix Figures B.1–B.7**). The 1D NMR spectra indicated the presence of six amide protons (δ_H 8.49, 7.93, 7.86, 7.58, 7.19, 6.76), seven carboxylic carbons (δ_C 174.0, 173.6, 173.4, 170.5,

169.9, 169.8, 159.8), and two ketone carbons (δ_C 206.2, 202.4). In addition, two α-protons (δ_H 4.32, 4.21) were detected, suggesting the presence of two amino acid residues. Based on the MS/MS data, these residues were proposed as threonine residues (**Figure 5.9**, **Appendix Table A.3**).

Two NH protons (δ_H 7.19, 6.76) showed HMBC correlations to a carbonyl carbon (δ_C 173.6) and a methylene carbon (δ_C 39.9), suggesting the presence of a terminal amide group. COSY and HMBC correlations completed a terminal 4-amino-3,5-dihydroxypentanamide (Ahpa) moiety as the first complete spin system from these resonances. COSY and HMBC correlations established two additional spin systems, revealing threonine amino acid residues (^{1}Thr and ^{2}Thr). The ^{1}H-NMR spectra showed the presence of four vinylic protons (δ_H 5.36 – 5.35), which exhibited COSY correlations to four allylic methylenes (δ_H 1.99 – 1.90). Continued COSY and HMBC correlations revealed a 3-amido-5,19-dihydroxyicosa-10,14-dienamide (Adhda) moiety as the fourth completed spin system. From a ketone carbon at δ_C 206.2, HMBC and COSY correlations were used to assign a fifth spin system as a 2-hydroxy-4-oxoheptanedioic acid (Hoha) functionality. The sixth spin system was completed based on HMBC and COSY correlations from an additional downfield ketone carbon at δ_C 202.4 to give a 2-methyl-propan-1-one (Mpo) moiety. The arrangement of the various spin systems was assigned based on HMBC correlations between the amide carbon of ^{1}Thr-1 and Ahpa-NH, ^{1}Thr-NH and ^{2}Thr-1, ^{2}Thr-NH and Adhda-1, Adhda-19 and Hoha-2 (**Figure 5.10**). The complete planar structure could not be determined due to unobserved HMBC correlations from Mpo-1 to the main structure. Potential connections were considered in the carbonyl groups of Adhda-21, Hoha-1, or Hoha-7.

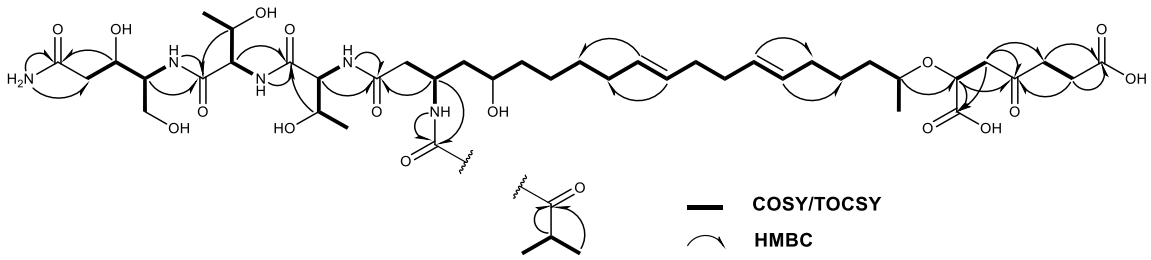


Figure 5.10 Key COSY/TOCSY and HMBC Correlations for Megapolipeptin A (5.2).

Compound **5.2** was treated with trimethylsilyl (TMS) diazomethane to determine possible connection points of the Mpo moiety and unambiguously assign the carboxylic acid function groups. MS analysis of the methylated product exhibited two peaks, each with an increase in mass of +42 Da ($[M+H]^+$ *m/z* 1000.5632 (calc'd *m/z* 1000.5670) $C_{48}H_{82}N_5O_{17}$), suggesting the methylation of three carboxylic acids (**Appendix Figures A.18 and A.19**). The derivatized products (designated as **5.4** and **5.5**) were isolated and subjected to NMR analysis. The ^{1}H-NMR spectrum of both products showed two methoxy signals assigned to the acid moieties of Hoha-1 and Hoha-7. Closer inspection of the ^{13}C-NMR spectrum revealed that the carbonyl carbon Mpo-1 was no longer present. Instead, COSY correlations showed this moiety was now a 2-isopropyl oxirane (Ipo) due to a Büchner-Curtius-Schlotterbeck reaction between the carbonyl group and TMS diazomethane to generate an epoxide (**Figure 5.11**). HMBC correlations from Ipo-2 (δ_H 2.27) and Ipo-5 (δ_H 2.87, 2.60) to Adhda-21 (δ_C 168.4) in both derivatized products established the connection of the Ipo moiety with the rest of the molecule (**Figure 5.11**). This completed the planar structures of **5.4** and **5.5** (**Figure 5.12**). The Mpo moiety was found to be connected to Adhda-21 in the unreacted NP, thereby completing the planar structure of megapolipeptin A (**5.2**) (**Figure 5.8**). Positional assignments of Ahpa [*m/z* 149.09352], ^{1}Thr [*m/z* 250.13945], and ^{2}Thr [*m/z* 351.19006] spin system moieties were corroborated by MS/MS fragmentation data (**Figure 5.9**).

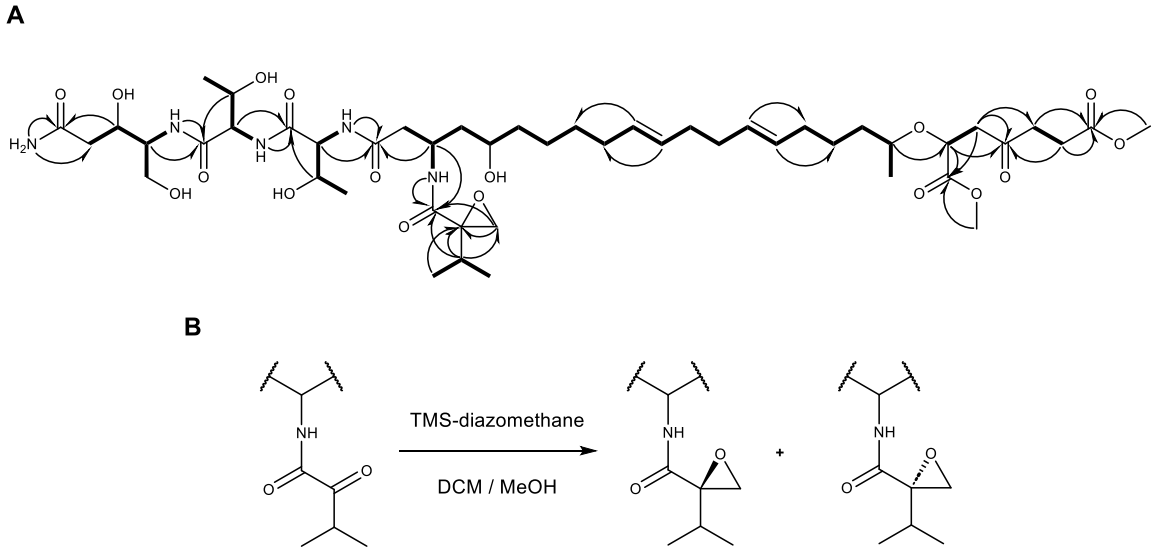


Figure 5.11 Key COSY and HMBC Correlations for the Derivatized NP with TMS Diazomethane.
(**A**) The key COSY and HMBC correlations for the derivatized megapolipeptin A (**5.2**) molecule display the connection of the Mpo moiety to the molecule. (**B**) The

proposed pathway via the Büchner-Curtius-Schlotterbeck rearrangement that converted a ketone to an epoxide in compounds **5.4** and **5.5**.

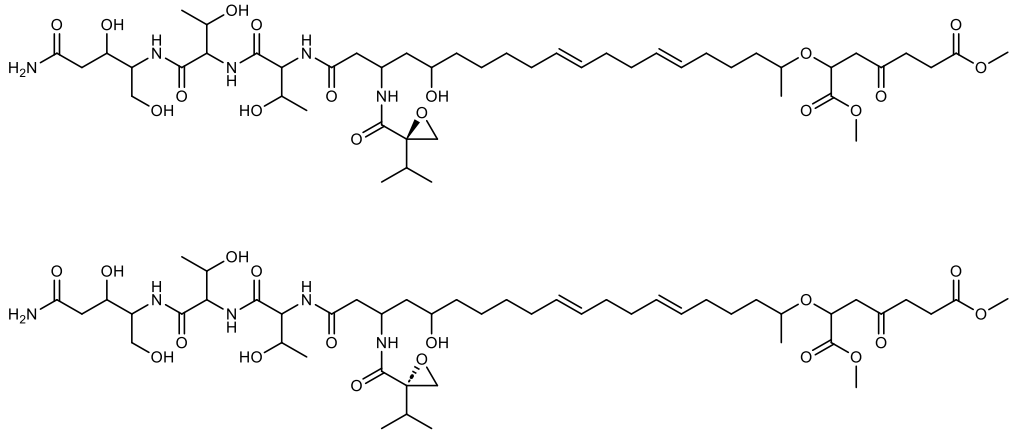


Figure 5.12 Planar Structures of the Megapolipeptin A Methyl Ester Derivatives (5.4 and 5.5).

Megapolipeptin B (**5.3**), also isolated as a white powder, exhibited a protonated molecular ion $[M+H]^+$ *m/z* 984.53821 (calc'd *m/z* 984.53872, -0.518 ppm) during HR-MS experiments, indicating a molecular formula of $C_{47}H_{77}N_5O_{17}$ (**Appendix Figure A.20**). A comparison of the MS spectra between compounds **5.2** and **5.3** revealed a mass difference of +26 Da, suggesting compound **5.3** contains two additional vinylic methines along the hydrocarbon chain (**Figure 5.13**).

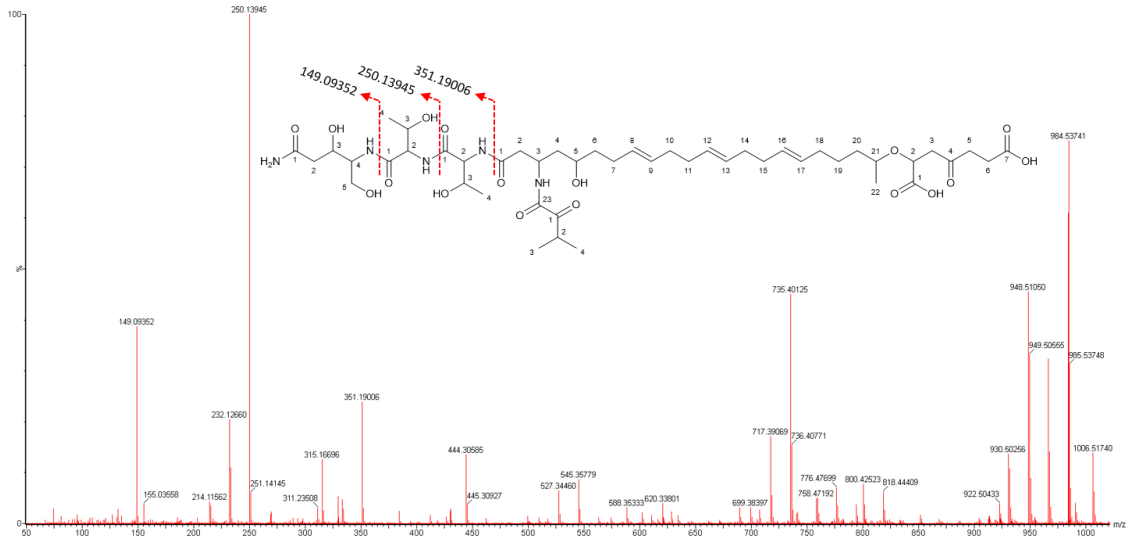


Figure 5.13 MS/MS Data for Megapolipeptin B (5.3).

The NMR data of **5.3** closely resembled those of **5.2**, showcasing six distinct spin systems (**Appendix Figures B.8 – B.13**). Upon thorough examination of the downfield region of the ^{1}H-NMR spectrum, containing the vinylic signals (δ_H 5.37), integration revealed the presence of six protons, consistent with the addition of 26 mass units. Comparative analysis of the 1D and 2D NMR spectra with compound **5.2** identified Ahpa, ^{1}Thr, ^{2}Thr, Hoha, and Mpo moieties. COSY and HMBC experiments established the final spin system as 3-amido-5,21-dihydroxydocosa-8,12,16-trienamide (Adhta) (**Appendix Table A.4**, **Figure 5.14**).

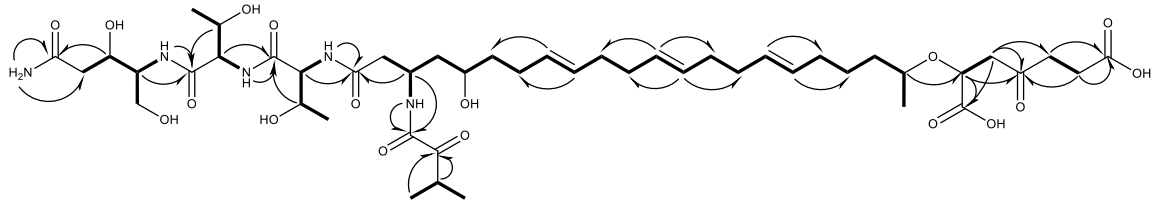


Figure 5.14 Key COSY and HMBC Correlation to Establish the Planar Structure of Megapolipeptin B (5.3).

The assignment of the Mpa moiety was determined similarly to **5.2**. Compound **5.3** was treated with TMS diazomethane to generate the 2-isopropyloxirane containing diastereomers (**5.6** and **5.7**) ([M+H]$^{+}$ 1026.5837 (calc'd 1026.5862) (**Appendix Figures A.21** and **A.22**). The complete planar structure of the derivatized product was established through COSY and HMBC correlations (**Figure 5.15**). Specifically, an HMBC correlation was observed between Mpo-1 and Adhta-23, suggesting that the Mpo moiety in **5.3** is connected to Adhta-21. Finally, the positions of the Ahpa [*m/z* 149.09352], ^{1}Thr [*m/z* 250.13945], and ^{2}Thr [*m/z* 351.19006] moieties were confirmed by MS/MS (**Figure 5.13**).

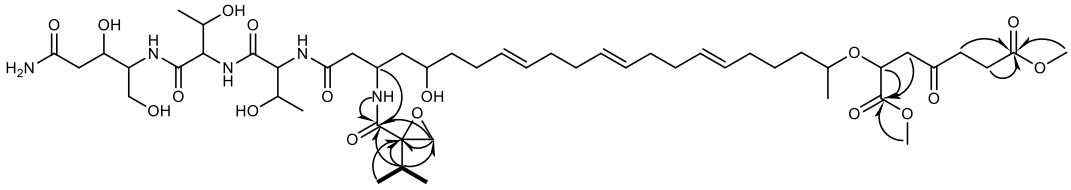


Figure 5.15 COSY and HMBC Correlations to Establish the Planar Structure of the Derivatized Methyl Ester Product (5.6 and 5.7).

Upon completing the planar structures of **5.2** and **5.3**, attention was directed at ascertaining the absolute configuration of the ^{1}Thr and ^{2}Thr residues in both molecules. The molecules underwent acid hydrolysis, liberating the amino acid hydrolysates.

Subsequently, these liberated hydrolysates were derivatized using Marfey's reagent (1-fluoro-2,4-dinitrophenyl-5-L-valineamide) to form diastereomers (**Section 5.5.2**).[181] Amino acid standards were prepared in parallel. The derivatized hydrolysates and amino acid standards were then analyzed via UPLC-MS. By comparing the retention times and *m/z* values of the derivatized hydrolysates to the standards, we determined the identity of the residues in both molecules. The analysis revealed the presence of L-threonine and L-*allo*-threonine (**Figure 5.16**).

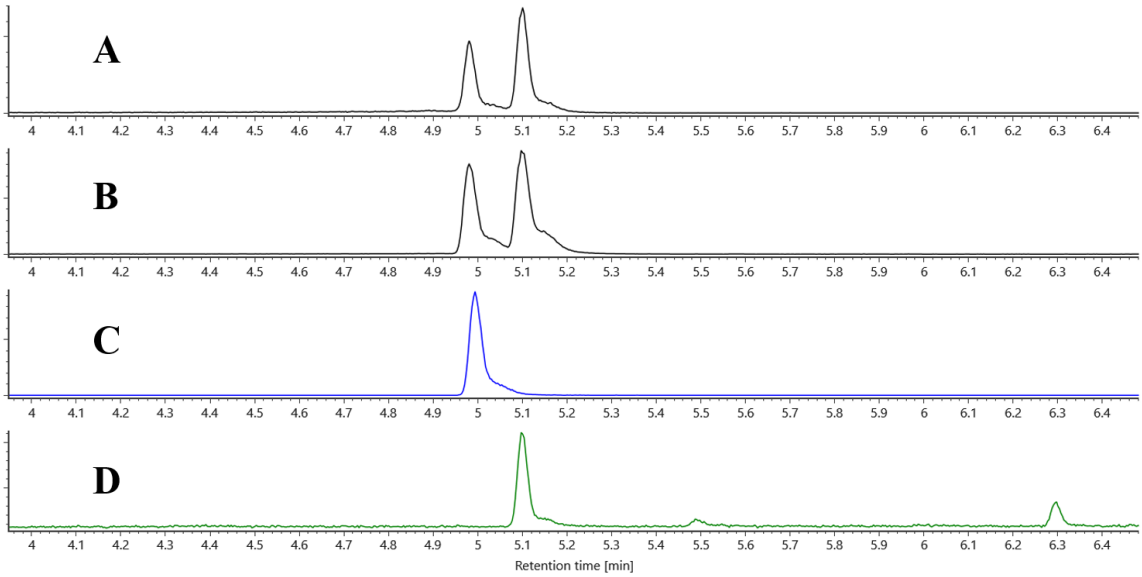


Figure 5.16 Marfey's Analysis of Megapolipeptin A (5.2) and Megapolipeptin B (5.3).
(**A**) EIC $[M+H]^+$ *m/z* 400.15 of megapolipeptin A hydrolysate. (**B**) EIC $[M+H]^+$ *m/z* 400.15 of megapolipeptin B hydrolysate. (**C**) FDVA-L-threonine standard. (**D**) FDVA-L-*allo*-threonine standard.

To determine the positions of the two different threonine residues, compound **5.3** underwent partial hydrolysis to obtain a molecule fragment containing only one threonine. Once a hydrolysis product containing a single threonine is identified, it can be isolated and subjected to Marfey's analysis to determine the absolute configuration of the threonine residue. This analysis allows the inference of the other threonine at the second position in the molecule. Initially, a small aliquot of compound **5.3** was partially hydrolyzed using 1N HCl at 110 °C over multiple time points to optimize the production of the desired fragment. MS analysis of the time points revealed a reaction product with an $[M+H]^+$ *m/z* 753.4213 at 3.83 minutes. This peak corresponded to the loss of the peptidic end of the molecule while retaining one threonine moiety (**Figure 5.17**). Notably,

this fragment was most abundant at the 30-minute time point. A larger sample of **5.3** was further hydrolyzed using the 30-minute target. After quenching the reaction and concentrating the mixture, the fragment was isolated by HPLC (**see Section 5.5.3**). The isolated fragment then underwent Marfey's analysis, confirming the associated amino acid as L-threonine (**Figure 5.18**). This inference allowed us to identify the other threonine residue as L-*allo*-threonine and assign the absolute configuration of both amino acid moieties in each molecule.

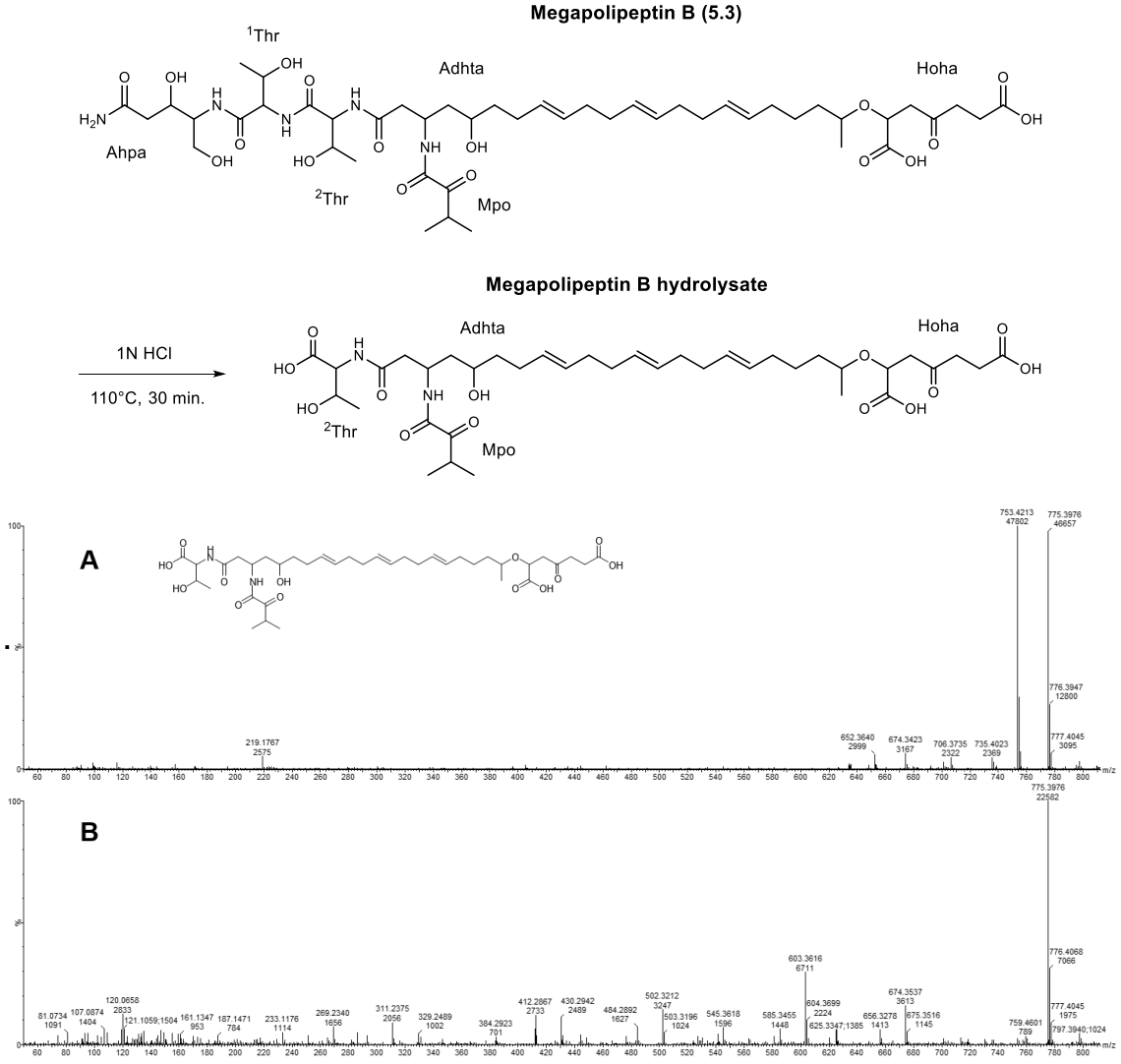


Figure 5.17 Hydrolysate Product Following Partial Hydrolysis of Megapolipeptin B (5.3).
The MS and MS/MS spectra were obtained for the hydrolysate product following partial hydrolysis of megapolipeptin B (**5.3**). The observed product peak (*m/z* 753.4213) corresponds to removing the Ahpa and [1]Thr moieties, leaving a single threonine residue in the isolated fragment. The peak was targeted for HPLC

isolation and subsequent Marfey's analysis for absolute configuration determination.

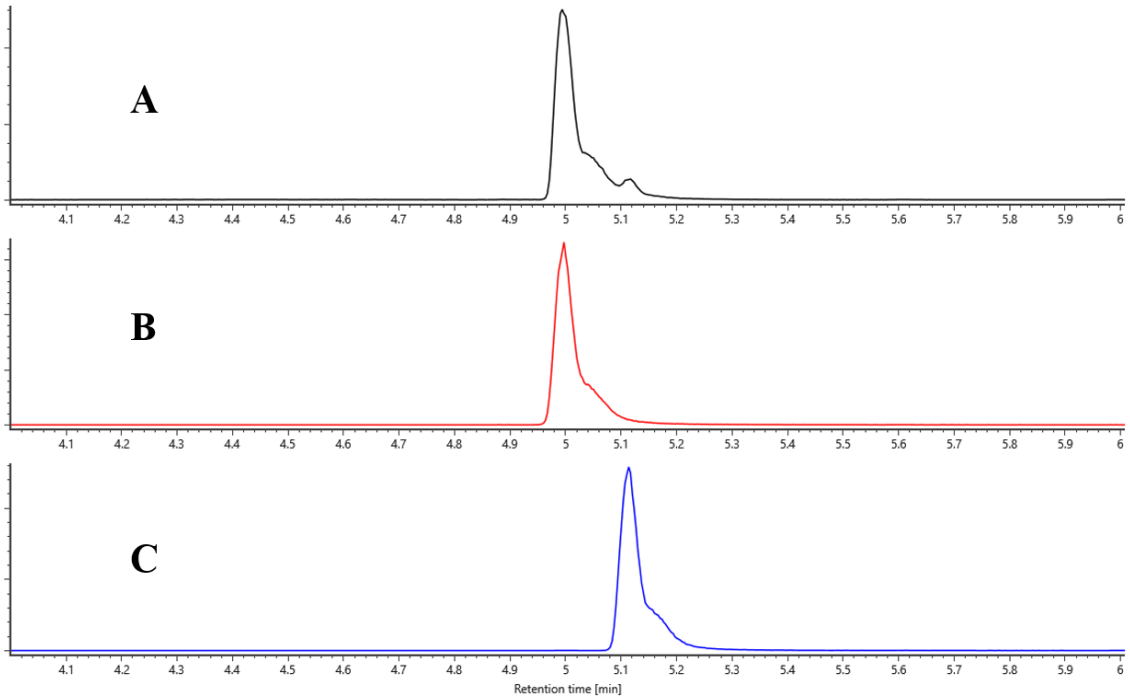


Figure 5.18 Marfey's Analysis on the Partial Hydrolysis Product. (**A**) EIC $[M+H]^+$ *m/z* 400.15 of megapolipeptin B (**5.3**) hydrolysate following partial hydrolysis and Marfey's analysis. (**B**) EIC $[M+H]^+$ *m/z* 400.15 of the FDVA-L-threonine standard. (**C**) EIC $[M+H]^+$ *m/z* 400.15 of the FDVA-L-*allo*-threonine standard.

Megapolipeptin A (5.2)

$[\alpha]_D^{20}$ +5.0 (c 0.23, MeOH); UV (MeOH) λmax (log ε) = 248 (1.41) nm. See **Appendix Table A.3** for NMR shifts. See **Appendix Figures B.1 – B.7** for NMR spectra. HR-ESI-MS $[M+H]^+$ *m/z* 958.52350 (calc'd for $C_{45}H_{75}N_5O_{17}$, 958.52307, 0.449 ppm).

Megapolipeptin B (5.3)

$[\alpha]_D^{20}$ +4.0 (c 0.52, MeOH); UV (MeOH) λmax (log ε) = 248 (2.04) nm. See **Appendix Table A.4** for NMR shifts. See **Appendix Figures B.8 – B.13** for NMR spectra. HR-ESI-MS $[M+H]^+$ *m/z* 984.53821 (calc'd for $C_{47}H_{77}N_5O_{17}$, 984.53872, -0.518 ppm).

5.3.3. Bioactivity Evaluation of Isolated Molecules

Previously, compounds **5.2** and **5.3** were predicted from a silent BGC. These large molecules likely serve a specific purpose within the organism, as their synthesis would otherwise be energetically costly. We evaluated the antimicrobial potential of the compounds by subjecting them to various tests. First, compounds **5.2** and **5.3** were evaluated for antibacterial susceptibility against a panel of 17 bacterial pathogens using our BioMAP antibacterial platform (**Appendix Tables A.5** and **A.6**).[101] No growth inhibition of pathogenic organisms was observed even at a maximum concentration of 128 µM (**Appendix Table A.5**). Furthermore, compounds **5.2** and **5.3** showed no antimycotic activity against *Candida albicans* or *Saccharomyces cerevisiae* even at a top concentration of 100 µM (**Appendix Table A.7** and **A.8**). Also, neither compound exhibited activity against *Aspergillus niger* or *Purpureocillium lilacinum* in qualitative filamentous screening assays, up to a maximum concentration of 100 µM (**Appendix Table A.7** and **A.8**). For detailed information on the antimicrobial experiments, refer to **Section 5.5.4**.

5.4. Conclusion

The studies discussed in this chapter aimed to leverage recent advancements in multi-omic applications to discover novel secondary metabolite NPs from Burkholderiales bacteria. These approaches were designed to be robust methods applicable to bacterial species beyond the Burkholderiales order. The primary objective was to overcome the bottleneck of obtaining NPs predicted from BGCs using bioinformatic methods, thereby expanding the known chemical space of NPs. Notably, the *P. megapolitana* clade emerged as a prolific producer of bioactive NPs, exhibiting a high BGC count to genome size ratio and a large diversity of PK-NRPS BGCs, including a conserved *mgp* BGC.[146] The *Burkholderia* strain FERM BP-3421 was identified as a successful producer of NP pharmaceutical candidates.[168] These studies underscored the strain's utility in producing novel NPs and serving as a host for predicted NPs from silent BGCs in native strains.

In **Section 5.2**, the study emphasized the synergistic application of genomic, bioinformatic, and metabolomic techniques to overcome the challenge of translating predictions into NPs. These methods were employed on *Burkholderia* sp. FERM BP-

3421, leading to the identification of a novel octadecalipodepsipeptide called selethramide (**5.1**) (**Figure 5.2**). This sizable peptide comprises a repetitive sequence of hydrophilic amino acids (7 serines, 1 *allo*-threonine) and hydrophobic amino acids (10 leucines), along with an N-terminal (*R*)-3-hydroxybutyric acid. Despite its weak antimicrobial activity, selethramide plays a role in surface motility due to its surfactant properties. The analysis presented here highlights enhanced DNA transfer efficiency and opens avenues for discovering future NPs from the *Burkholderia* strain.

The second study (**Section 5.3**) presented the discovery of two novel NPs originating from a silent BGC within a *P. megapolitana* strain. Researchers focused on a hybrid PK-NRPS BGC in a *P. megapolitana* strain that is inherently silent in the bacterium. The BGC *mgp* was identified as a unique ensemble of enzymes critical to the biosynthesis of megapolipeptins through bioinformatics and genomics analyses. Although MS detected no NPs in the wild-type strain, bioinformatics predictions hinted at the presence of novel NPs with potential bioactivity. The *mgp* was cloned and heterologously expressed in the host strain *Burkholderia* sp. FERM BP-3421. This effort led to the discovery of megapolipeptins A (**5.2**) and B (**5.3**) at concentrations of 0.6 mg/L and 1.5 mg/L, respectively (**Figure 5.8**).

These studies have demonstrated that genome sequencing enhances our comprehension of biodiversity and facilitates the exploration of novel and biologically valuable NPs. The continued advancements in the multi-omic disciplines invigorate the search for secondary metabolites in unexplored realms, uncovering a wealth of potential.[162] One of the described studies employed multidisciplinary approaches to obtain NPs from silent or underexpressed sources, mainly when the isolated molecules exhibit chemical diversity and low structural similarity to reported microbial NPs. The genomics-driven and synthetic biology-enabled pipeline described herein is invaluable for expanding the chemical space by discovering novel NPs.

5.5. Materials and Methods

5.5.1. General Experimental Procedures

Optical rotations were measured on a Model 341 (Perkin Elmer) polarimeter. Ultraviolet absorption spectra were recorded on a Shimadzu UV-3600 Plus UV–VIS

spectrophotometer. HR-MS and MS/MS experiments were recorded on a Waters Select Series MRT and a SYNAPT G2-Si UPLC-ESI-qTOF mass spectrometers. NMR spectra were measured on an AVANCE II 600 MHz spectrometer equipped with a 5 mm QCI cryoprobe and referenced to residual solvent proton and carbon signals δ_H 2.50 and δ_C 39.5 for DMSO-*d*$_6$. MPLC (CombiFlash, Teledyne ISCO) was performed on a RediSep Rf solid load cartridge (5g, Teledyne ISCO). HPLC separations were performed on an Agilent 1200 series equipped with a binary pump and a diode array detector using a Phenomenex Synergi-RP FUSION (250 x 10 mm, 10 µm), a Phenomenex Kinetex XB-C18 (250 x 4.6 mm, 5 µm), and a VyDAC-RP (250 mm x 4.6 mm, 5 µm) columns. Solvents used for HPLC chromatography were optima grade and were used without further purification.

5.5.2. Determination of the Absolute Configuration of Amino Acids

Isolated compounds (300 µg each) were hydrolyzed (100 µL 6N HCl at 110 °C for 18 hours) in a 500 µL sealed reaction vessel. After the reaction, the excess HCl was removed under a stream of N_2 (g), and the residue was re-suspended in 100 µL 50% aq. acetone. 100 µL of FDVA solution in acetone (10 mg/mL) and 40 µL of 1M $NaHCO_3$ were added to the reaction vessel and heated at 40 °C for 1 hour. The reaction was then quenched with 40 µL 1N HCl and dried under a N_2 (g). The dried residue was dissolved in 500 µL MeOH and compared with standard FDVA-amino acids by UHPLC-MS (Waters ACQUITY HSS T3 1.8 µm, 0.650 mL/min) using a gradient of 5% aq. ACN to 95% aq. ACN with 0.02% formic acid over 13 minutes. The absolute configuration of each amino acid was determined by comparing retention times of FDVA-amino acid derivatives in target compounds with standard D,L FDVA-amino acids.

5.5.3. Partial Hydrolysis and Isolation of the Hydrolysate

Megapolipeptin B (**5.3**, 0.5 mg) underwent partial hydrolysis to isolate a molecule fragment containing one threonine residue. Compound **5.3** was hydrolyzed using 1N HCl at 110 °C for 30 minutes in a sealed 500 µL reaction vessel. MS continuously monitored the reaction progress to optimize the yield of the desired product. After 30 minutes, the reaction was quenched with 1M $NaHCO_3$ and dried under a stream of N_2 (g). The resulting hydrolysate was then purified using HPLC on an Agilent 1200 series system equipped with a Kinetex XB-C_{18} Phenomenex column (250 x 4.6 mm, 5 µm). An elution

gradient from 5% to 95% aq. ACN over 40 minutes was employed, with a 1.25 mL/min flow rate. The compound eluted at 25.1 minutes (0.1 mg) and was monitored by UV/Vis at λ 254 nm.

5.5.4. Antimicrobial Screening

Antimicrobial susceptibility tests for target compounds against the bacterial target panel are performed using a miniaturized HT assay adapted from the broth microdilution method outlined by the Clinical and Laboratory Standards Institute (CLSI). Bacterial test strains were individually grown on fresh Nutrient Broth (NB, ATCC Medium 3) agar, Tryptic Soy Broth (TSB, ATCC Medium 18) agar, or Brain Heart Infusion (BHI, ATCC Medium 44) agar, respectively (**Appendix Table A.5**), as recommended by the American Type Culture Collection (ATCC) cultivation protocol. Individual colonies were used to inoculate 3 mL of sterile NB, TSB, or BHI media and grown overnight with shaking (200 RPM; 37 °C). *Listeria ivanovii* (ATCC BAA-139) was incubated overnight but not shaken (37 °C; 5% CO_2). Saturated overnight cultures were diluted in their respective media according to turbidity to achieve approximately 5 x 105 CFU/mL of final inoculum density and dispensed into sterile, clear polystyrene 384-well microplates (Thermo Scientific™ 265202) with a final screening volume of 30 µL. *L. ivanovii* was diluted and grown in Haemophilus Test Medium (HTM; ATCC Medium 2167). DMSO solutions of test compounds and antibiotic controls were prepared as a 1:1 dilution series and pinned into each assay plate (200 nL) using a high throughput pinning robot (Tecan Freedom EVO 100) to achieve final screening concentrations ranging from 128 µM to 3.91 nM per compound. In each 384-well plate, lane one was reserved for the DMSO vehicle and culture medium; lane two was reserved for the DMSO vehicle, culture medium, and target bacteria; lanes 23 and 24 were reserved for antibiotic controls, DMSO vehicle, culture medium, and target bacteria. After compound pinning, assay plates were read as t_0 at OD_{600} using an automated plate reader (BioTek Synergy Neo2), sealed with a lid, and placed in a humidity-controlled incubator at 37 °C for 18-20 hours before optical density was obtained for t_{20}. *L. ivanovii* was incubated in a separate incubator (37 °C; 5% CO_2). The resulting growth curves for each dilution series were used to determine the MIC values for each tested compound following standard procedures.

References

1. Mata, J. M.; van der Nol, E.; Pomplun, S. J. Advances in Ultrahigh Throughput Hit Discovery with Tandem Mass Spectrometry Encoded Libraries. *J. Am. Chem. Soc.* **2023**, *145* (34), 19129–9139. DOI: 10.1021/jacs.3c04899.

2. Rütten, A.; Kirchner, T.; Musiol-Kroll, E. M. Overview on Strategies and Assays for Antibiotic Discovery. *Pharmaceuticals.* **2022**, *15* (10), 1302–1326. DOI: 10.3390/ph15101302.

3. Vincent, F.; Nueda, A.; Lee, J.; Schenone, M.; Prunotto, M.; Mercola, M. Phenotypic Drug Discovery: Recent Successes, Lessons Learned and New Directions. *Na.t Rev. Drug Discov.* **2022**, *21* (12), 899–914. DOI: 10.1038/s41573-022-00472-w.

4. Vitale, G. A.; Geibel, C.; Minda, V.; Wang, M.; Aron, A. T.; Petras, D. Connecting Metabolome and Phenotype: Recent Advances in Functional Metabolomics Tools for the Identification of Bioactive Natural Products. *Nat. Prod. Rep.* **2024**, *41* (6), 885–904. DOI: 10.1039/d3np00050h.

5. Plumb, R. S.; Gethings, L. A.; Rainville, P. D.; Isaac, G.; Trengove, R.; King, A. M.; Wilson, I. D. Advances in High Throughput LC/MS Based Metabolomics: A Review. *Trends Anal. Chem.* **2023**, *160*, 116954. DOI: 10.1016/j.trac.2023.116954.

6. Atanasov, A. G.; Zotchev, S. B.; Dirsch, V. M.; Supuran, C. T. Natural Products in Drug Discovery: Advances and Opportunities. *Nat. Rev. Drug Discov.* **2021**, *20* (3), 200–216. DOI: 10.1038/s41573-020-00114-z.

7. Gavriilidou, A. F. M.; Sokratous, K.; Yen, H.-Y.; De Colibus, L. High-Throughput Native Mass Spectrometry Screening in Drug Discovery. *Front. Mol. Biosci.* **2022**, *9*, 837901. DOI: 10.3389/fmolb.2022.837901.

8. Corson, T. W.; Crews, C. M. Molecular Understanding and Modern Application of Traditional Medicines: Triumphs and Trials. *Cell* **2007**, *130* (5), 769–774. DOI: 10.1016/j.cell.2007.08.021.

9. Moffat, J. G.; Vincent, F.; Lee, J. A.; Eder, J.; Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nat. Rev. Drug Discov.* **2017**, *16* (8), 531–543. DOI: 10.1038/nrd.2017.111.

10. Dueñas, M. E.; Peltier-Heap, R. E.; Leveridge, M.; Annan, R. S.; Büttner, F. H.; Trost, M. Advances in High-throughput Mass Spectrometry in Drug Discovery. *EMBO Mol. Med.* **2023**, *15*, e14850. DOI: 10.15252/emmm.202114850.

11. Jiang, L.-X.; Yang, M.; Wali, S. N.; Laskin, J. High-Throughput Mass Spectrometry Imaging of Biological Systems: Current Approaches and Future Directions. *Trends Anal. Chem.* **2023**, *163*, 117055. DOI: 10.1016/j.trac.2023.117055A.

12. Gaudêncio, S. P.; Bayram, E.; Lukić Bilela, L.; Cueto, M.; Díaz-Marrero, A. R.; Haznedaroglu, B. Z.; Jimenez, C.; Mandalakis, M.; Pereira, F.; Reyes, F.; Tasdemir, D. Advanced Methods for Natural Products Discovery: Bioactivity Screening, Dereplication, Metabolomics Profiling, Genomic Sequencing, Databases and Informatic Tools, and Structure Elucidation. *Mar. Drugs.* **2023**, *21* (5), 308. DOI: 10.3390/md21050308.

13. Priessner, M.; Lewis, R. J.; Johansson, M. J.; Goodman, J. M.; Janet, J. P.; Tomberg, A. HSQC Spectra Simulation and Matching for Molecular Identification. *J. Chem. Inf. Model.* **2024**, *64* (8), 3180–3191. DOI: 10.1021/acs.jcim.3c01735.

14. Abram, K. J.; McCloskey, D. A Comprehensive Evaluation of Metabolomics Data Preprocessing Methods for Deep Learning. *Metabolites.* **2022**, *12* (3), 202. DOI: 10.3390/metabo12030202.

15. Ma, X. Recent Advances in Mass Spectrometry-Based Structural Elucidation Techniques. *Molecules.* **2022**, *27* (19), 6466. DOI: 10.3390/molecules27196466.

16. Hight, S. K.; Clark, T. N.; Kurita, K. L.; McMillan, E. A.; Bray, W.; Shaikh, A. F.; Khadilkar, A.; Haeckl, F. P. J.; Carnevale-Neto, F.; La, S.; Lohith, A.; Vaden, R. M.; Lee, J.; Wei, S.; Lokey, R. S.; White, M. A.; Linington, R. G.; MacMillan, J. B. High-Throughput Functional Annotation of Natural Products by Integrated Activity Profiling. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (49), e2208458119. DOI: 10.1073/pnas.2208458119.

17. Ayon, N. J. High-Throughput Screening of Natural Product and Synthetic Molecule Libraries for Antibacterial Drug Discovery. *Metabolites.* **2023**, *13* (5), 625. DOI: 10.3390/metabo13050625.

18. Wilson, B. A. P.; Thornburg, C. C.; Henrich, C. J.; Grkovic, T.; O'Keefe, B. R. Creating and Screening Natural Product Libraries. *Nat. Prod. Rep.* **2020**, *37* (7), 893–918. DOI: 10.1039/c9np00068b.

19. Grkovic, T.; Akee, R. K.; Thornburg, C. C.; Trinh, S. K.; Britt, J. R.; Harris, M. J.; Evans, J. R.; Kang, U.; Ensel, S.; Henrich, C. J.; Gustafson, K. R.; Schneider, J. P.; O'Keefe, B. R. National Cancer Institute (NCI) Program for Natural Products Discovery: Rapid Isolation and Identification of Biologically Active Natural Products from the NCI Prefractionated Library. *ACS Chem. Biol.* **2020**, *15* (4), 1104–1114. DOI: 10.1021/acschembio.0c00139

20. Thornburg, C. C.; Britt, J. R.; Evans, J. R.; Akee, R. K.; Whitt, J. A.; Trinh, S. K.; Harris, M. J.; Thompson, J. R.; Ewing, T. L.; Shipley, S. M.; Grothaus, P. G.; Newman, D. J.; Schneider, J. P.; Grkovic, T.; O'Keefe, B. R. NCI Program for Natural Product Discovery: A Publicly-Accessible Library of Natural Product Fractions for High-Throughput Screening. *ACS Chem. Biol.* **2018**, *13* (9), 2484–2497. DOI: 10.1021/acschembio.8b00389.

21. Camp, D.; Campitelli, M.; Carroll, A. R.; Davis, R. A.; Quinn, R. J. Front-Loading Natural-Product-Screening Libraries for Log P: Background, Development, and Implementation. *Chem. Biodivers.* **2013**, *10* (4), 524–537. DOI: 10.1002/CBDV.201200302.

22. Appleton, D. R.; Buss, A. D.; Butler, M. S. A Simple Method for High-Throughput Extract Prefractionation for Biological Screening. *Chimia (Aarau).* **2007**, *61* (6), 327. DOI: 10.2533/chimia.2007.327.

23. Hartl, D.; de Luca, V.; Kostikova, A.; Laramie, J.; Kennedy, S.; Ferrero, E.; Siegel, R.; Fink, M.; Ahmed, S.; Millholland, J.; Schuhmacher, A.; Hinder, M.; Piali, L.; Roth, A. Translational Precision Medicine: An Industry Perspective. *J. Transl. Med.* **2021**, *19* (1), 245. DOI: 10.1186/s12967-021-02910-6.

24. Mullowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; Durairaj, J.; Gorostiola González, M.; Helfrich, E. J. N.; Huber, F.; Leopold-Messer, S.; Rajan, K.; de Rond, T.; van Santen, J. A.; Sorokina, M.; Balunas, M. J.; Beniddir, M. A.; van Bergeijk, D. A.; Carroll, L. M.; Clark, C. M.; Clevert, D.-A.; Dejong, C. A.; Du, C.; Ferrinho, S.; Grisoni, F.; Hofstetter, A.; Jespers, W.; Kalinina, O. V.; Kautsar, S. A.; Kim, H.; Leao, T. F.; Masschelein, J.; Rees, E. R.; Reher, R.; Reker, D.; Schwaller, P.; Segler, M.; Skinnider, M. A.; Walker, A. S.; Willighagen, E. L.; Zdrazil, B.; Ziemert, N.; Goss, R. J. M.; Guyomard, P.; Volkamer, A.; Gerwick, W. H.; Kim, H. U.; Müller, R.; van Wezel, G. P.; van Westen, G. J. P.; Hirsch, A. K. H.; Linington, R. G.; Robinson, S. L.; Medema, M. H. Artificial Intelligence for Natural Drug Discovery. *Nat. Rev. Drug Discov.* **2023**, *22* (11), 895–916. DOI: 10.1038/s41573-023-00774-7.

25. Caesar, L. K.; Montaser, R.; Keller, N. P.; Kelleher, N. L. Metabolomics and Genomics in Natural Products Research: Complementary Tools for Targeting New Chemical Entities. *Nat. Prod. Rep.* **2021**, *38* (11), 2041–2065. DOI: 10.1039/D1NP00036E.

26. Heinrich, M.; Jalil, B.; Abdel-Tawab, M.; Echeverria, J.; Kulić, Ž.; McGaw, L. J.; Pezzuto, J. M.; Potterat, O.; Wang, J.-B. Best Practice in the Chemical Characterisation of Extracts Used in Pharmacological and Toxicological Research—The ConPhyMP—Guidelines. *Front. Pharmacol.* **2022**, *13*, 953205. DOI: 10.3389/fphar.2022.953205.

27. Lee, S.; van Santen, J. A.; Farzaneh, N.; Liu, D. Y.; Pye, C. R.; Baumeister, T. U. H.; Wong, W. R.; Linington, R. G. NP Analyst: An Open Online Platform for Compound Activity Mapping. *ACS Cent. Sci.* **2022**, *8* (2), 223–234. DOI: 10.1021/acscentsci.1c01108.

28. Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G. Retrospective Analysis of Natural Products Provides Insights for Future Discovery Trends. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (22), 5601–5606. DOI: 10.1073/pnas.1614680114.

29. Clark, T. N.; Houriet, J.; Vidar, W. S.; Kellogg, J. J.; Todd, D. A.; Cech, N. B.; Linington, R. G. Interlaboratory Comparison of Untargeted Mass Spectrometry Data Uncovers Underlying Causes for Variability. *J. Nat. Prod.* **2021**, *84* (3), 824–835. DOI: 10.1021/acs.jnatprod.0c01376.

30. Harrison, C. Patenting Natural Products Just Got Harder. *Nat. Biotechnol.* **2014**, *32* (5), 403–404. DOI: 10.1038/NBT0514-403A.

31. Paslay, J. W.; Morin, J. E.; Harrison, R. K. High Throughput Screening in the Twenty-First Century. In *Topics in Medicinal Chemistry*; Springer, Berlin, Heidelberg, 2009; Vol. 5, pp 25–83. DOI: 10.1007/7355_2009_6.

32. Hansel, C. S.; Plant, D. L.; Holdgate, G. A.; Collier, M. J.; Plant, H. Advancing Automation in High-Throughput Screening: Modular Unguarded Systems Enable Adaptable Drug Discovery. *Drug Discov. Today.* **2022**, *27* (8), 2051–2056. DOI: 10.1016/j.drudis.2022.03.010.

33. Bray, M.-A.; Singh, S.; Han, H.; Davis, C. T.; Borgeson, B.; Hartland, C.; Kost-Alimova, M.; Gustafsdottir, S. M.; Gibson, C. C.; Carpenter, A. E. Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes. *Nat. Protoc.* **2016**, *11* (9), 1757–1774. DOI: 10.1038/nprot.2016.105.

34. Pahl, A.; Sievers, S. The Cell Painting Assay as a Screening Tool for the Discovery of Bioactivities in New Chemical Matter. In *Methods in Molecular Biology (Clifton, N.J.)*; Methods Mol. Biol., 2019; Vol. 1888, pp 115–126. DOI: 10.1007/978-1-4939-8891-4_6.

35. Cimini, B. A.; Chandrasekaran, S. N.; Kost-Alimova, M.; Miller, L.; Goodale, A.; Fritchman, B.; Byrne, P.; Garg, S.; Jamali, N.; Logan, D. J.; Concannon, J. B.; Lardeau, C.-H.; Mouchet, E.; Singh, S.; Shafqat Abbasi, H.; Aspesi Jr., P.; Boyd, J. D.; Gilbert, T.; Gnutt, D.; Hariharan, S.; Hernandez, D.; Hormel, G.; Juhani, K.; Melanson, M.; Mervin, L. H.; Monteverde, T.; Pilling, J. E.; Skepner, A.; Swalley, S. E.; Vrcic, A.; Weisbart, E.; Williams, G.; Yu, S.; Zapiec, B.; Carpenter, A. E. Optimizing the Cell Painting Assay for Image-Based Profiling. *Nat. Protoc.* **2023**, *18*, 1981–2013. DOI: 10.1038/s41596-023-00840-9.

36. Chaleckis, R.; Meister, I.; Zhang, P.; Wheelock, C. E. Challenges, Progress and Promises of Metabolite Annotation for LC-MS-Based Metabolomics. *Curr. Opin. Biotechnol.* **2019**, *55*, 44–50. DOI: 10.1016/j.copbio.2018.07.010.

37. Perez de Souza, L.; Alseekh, S.; Scossa, F.; Fernie, A. R. Ultra-High-Performance Liquid Chromatography High-Resolution Mass Spectrometry Variants for Metabolomics Research. *Nat. Methods.* **2021**, *18* (7), 733–746. DOI: 10.1038/s41592-021-01116-4.

38. Williams, J. D.; Pu, F.; Sawicki, J. W.; Elsen, N. L. Ultra-High-Throughput Mass Spectrometry in Drug Discovery: Fundamentals and Recent Advances. *Expert Opin. Drug Discov.* **2024**, *19* (3), 291–301. DOI: 10.1080/17460441.2023.2293153.

39. Häbe, T. T.; Liu, C.; Covey, T. R.; Simon, R. P.; Reindl, W.; Büttner, F. H.; Winter, M.; Bischoff, D.; Luippold, A. H.; Runge, F. Ultrahigh-Throughput ESI-MS: Sampling Pushed to Six Samples per Second by Acoustic Ejection Mass Spectrometry. *Anal. Chem.* **2020**, *92* (18), 12242–12249. DOI: 10.1021/acs.analchem.0c01632.

40. Zhang, H.; Liu, C.; Hua, W.; Ghislain, L. P.; Liu, J.; Aschenbrenner, L.; Noell, S.; Dirico, K. J.; Lanyon, L. F.; Steppan, C. M.; West, M.; Arnold, D. W.; Covey, T. R.; Datwani, S. S.; Troutman, M. D. Acoustic Ejection Mass Spectrometry for High-Throughput Analysis. *Anal. Chem.* **2021**, *93* (31), 10850–10861. DOI: 10.1021/acs.analchem.1c01137.

41. Liu, C. Acoustic Ejection Mass Spectrometry: Fundamentals and Applications in High-Throughput Drug Discovery. *Expert Opin. Drug Discov.* **2022**, *17* (7), 775–787. DOI: 10.1080/17460441.2022.2084069.

42. Smith, M. J.; Ivanov, D. P.; Weber, R. J. M.; Wingfield, J.; Viant, M. R. Acoustic Mist Ionization Mass Spectrometry for Ultrahigh-Throughput Metabolomics Screening. *Anal. Chem.* **2021**, *93* (26), 9258–9266. DOI: 10.1021/acs.analchem.1c01616.

43. Silva, R.; Lopes, N. P.; Silva, D. B. Application of MALDI Mass Spectrometry in Natural Products Analysis. *Planta Medica*. Planta Med May 1, 2016, pp 671–689. DOI: 10.1055/s-0042-104800.

44. Radosevich, A. J.; Pu, F.; Chang-Yen, D.; Sawicki, J. W.; Talaty, N. N.; Elsen, N. L.; Williams, J. D.; Pan, J. Y. Ultra-High-Throughput Ambient MS: Direct Analysis at 22 Samples per Second by Infrared Matrix-Assisted Laser Desorption Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **2022**, *94* (12), 4913–4918. DOI: 10.1021/acs.analchem.1c04605.

45. Woo, H. K.; Northen, T. R.; Yanes, O.; Siuzdak, G. Nanostructure-Initiator Mass Spectrometry: A Protocol for Preparing and Applying NIMS Surfaces for High-Sensitivity Mass Analysis. *Nat. Protoc.* **2008**, *3* (8), 1341–1349. DOI: 10.1038/nprot.2008.110.

46. Prudent, R.; Annis, D. A.; Dandliker, P. J.; Ortholand, J.-Y.; Roche, D. Exploring New Targets and Chemical Space with Affinity Selection-Mass Spectrometry. *Nat. Rev. Chem.* **2020**, *5*, 62–71. DOI: 10.1038/s41570-020-00229-2.

47. Veach, B. T.; Mudalige, T. K.; Rye, P. RapidFire Mass Spectrometry with Enhanced Throughput as an Alternative to Liquid-Liquid Salt Assisted Extraction and LC/MS Analysis for Sulfonamides in Honey. *Anal. Chem.* **2017**, *89* (6), 3256–3260. DOI: 10.1021/ACS.ANALCHEM.6B04889.

48. Sawyer, W. S.; Srikumar, N.; Carver, J.; Chu, P. Y.; Shen, A.; Xu, A.; Williams, A. J.; Spiess, C.; Wu, C.; Liu, Y.; Tran, J. C. High-Throughput Antibody Screening from Complex Matrices Using Intact Protein Electrospray Mass Spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (18), 9851–9856. DOI: 10.1073/PNAS.1917383117.

49. Miller, V. P. SPE-MS Analysis of Absorption, Distribution, Metabolism and Excretion Assays: A Tool to Increase Throughput and Steamline Workflow. *Bioanalysis.* **2012**, *4* (9), 1111–1121. DOI: 10.4155/BIO.12.86.

50. Clausse, V.; Tao, D.; Debnath, S.; Fang, Y.; Tagad, H. D.; Wang, Y.; Sun, H.; LeClair, C. A.; Mazur, S. J.; Lane, K.; Shi, Z. D.; Vasalatiy, O.; Eells, R.; Baker, L. K.; Henderson, M. J.; Webb, M. R.; Shen, M.; Hall, M. D.; Appella, E.; Appella, D. H.; Coussens, N. P. Physiologically Relevant Orthogonal Assays for the Discovery of Small-Molecule Modulators of WIP1 Phosphatase in High-Throughput Screens. *J. Bio. Chem.* **2019**, *294* (46), 17654–17668. DOI: 10.1074/JBC.RA119.010201.

51. Jian, W.; Romm, M. V.; Edom, R. W.; Miller, V. P.; Lamarr, W. A.; Weng, N. Evaluation of a High-Throughput Online Solid Phase Extraction-Tandem Mass Spectrometry System for in Vivo Bioanalytical Studies. *Anal. Chem.* **2011**, *83* (21), 8259–8266. DOI: 10.1021/ac202017c.

52. Žuvela, P.; Skoczylas, M.; Liu, J. J.; Ba Czek, T.; Kaliszan, R.; Wong, M. W.; Buszewski, B. Column Characterization and Selection Systems in Reversed-Phase High-Performance Liquid Chromatography. *Chem. Rev.* **2019**, *119*, 3674–3729. DOI: 10.1021/acs.chemrev.8b00246.

53. Van Santen, J. A.; Poynton, E. F.; Iskakova, D.; Mcmann, E.; Alsup, T. A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H.; Mccadden, C. A.; Parra, J.; Soldatou, S.; Rudolf, J. D.; Janssen, E. M. L.; Duncan, K. R.; Linington, R. G. The Natural Products Atlas 2.0: A Database of Microbially-Derived Natural Products. *Nucleic Acids Res.* **2022**, *50* (D1), D1317–D1323. DOI: 10.1093/NAR/GKAB941.

54. Terlouw, B. R.; Blin, K.; Navarro-Muñoz, J. C.; Avalon, N. E.; Chevrette, M. G.; Egbert, S.; Lee, S.; Meijer, D.; Recchia, M. J. J.; Reitz, Z. L.; van Santen, J. A.; Selem-Mojica, N.; Tørring, T.; Zaroubi, L.; Alanjary, M.; Aleti, G.; Aguilar, C.; Al-Salihi, S. A. A.; Augustijn, H. E.; Avelar-Rivas, J. A.; Avitia-Domínguez, L. A.; Barona-Gómez, F.; Bernaldo-Agüero, J.; Bielinski, V. A.; Biermann, F.; Booth, T. J.; Carrion Bravo, V. J.; Castelo-Branco, R.; Chagas, F. O.; Cruz-Morales, P.; Du, C.; Duncan, K. R.; Gavriilidou, A.; Gayrard, D.; Gutiérrez-García, K.; Haslinger, K.; Helfrich, E. J. N.; van der Hooft, J. J. J.; Jati, A. P.; Kalkreuter, E.; Kalyvas, N.; Kang, K. Bin; Kautsar, S.; Kim, W.; Kunjapur, A. M.; Li, Y. X.; Lin, G. M.; Loureiro, C.; Louwen, J. J. R.; Louwen, N. L. L.; Lund, G.; Parra, J.; Philmus, B.; Pourmohsenin, B.; Pronk, L. J. U.; Rego, A.; Rex, D. A. B.; Robinson, S.; Rosas-Becerra, L. R.; Roxborough, E. T.; Schorn, M. A.; Scobie, D. J.; Singh, K. S.; Sokolova, N.; Tang, X.; Udwary, D.; Vigneshwari, A.; Vind, K.; Vromans, S. P. J. M.; Waschulin, V.; Williams, S. E.; Winter, J. M.; Witte, T. E.; Xie, H.; Yang, D.; Yu, J.; Zdouc, M.; Zhong, Z.; Collemare, J.; Linington, R. G.; Weber, T.; Medema, M. H. MIBiG 3.0: A Community-Driven Effort to Annotate Experimentally Validated Biosynthetic Gene Clusters. *Nucleic Acids Res.* **2023**, *51* (D1), D603–D610. DOI: 10.1093/NAR/GKAC1049.

55. Zhang, C.; Idelbayev, Y.; Roberts, N.; Tao, Y.; Nannapaneni, Y.; Duggan, B. M.; Min, J.; Lin, E. C.; Gerwick, E. C.; Cottrell, G. W.; Gerwick, W. H. Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Sci. Rep.* **2017**, *7* (14243), 1–17. DOI: 10.1038/s41598-017-13923-x.

56. Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L. F.; Reher, R.; Kang, K. Bin; Van Der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **2021**, *84* (11), 2795–2807. DOI: 10.1021/acs.jnatprod.1c00399.

57. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P., C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P. M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34* (8), 828–837. DOI: 10.1038/nbt.3597.

58. Wang, F.; Allen, D.; Tian, S.; Oler, E.; Gautam, V.; Greiner, R.; Metz, T. O.; Wishart, D. S. CFM-ID 4.0 – a Web Server for Accurate MS-Based Metabolite Identification. *Nucleic Acids Res.* **2022**, *50* (W1), W165–W174. DOI: 10.1093/nar/gkac383.

59. Guijas, C.; Montenegro-Burke, J. R.; Domingo-Almenara, X.; Palermo, A.; Warth, B.; Hermann, G.; Koellensperger, G.; Huan, T.; Uritboonthai, W.; Aisporna, A. E.; Wolan, D. W.; Spilker, M. E.; Benton, H. P.; Siuzdak, G. METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Anal. Chem.* **2018**, *90* (5), 3156–3164. DOI: 10.1021/acs.analchem.7b04424.

60. Wishart, D. S.; Sayeeda, Z.; Budinski, Z.; Guo, A.; Lee, B. L.; Berjanskii, M.; Rout, M.; Peters, H.; Dizon, R.; Mah, R.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Varshavi, D.; Varshavi, D.; Oler, E.; Allen, D.; Cao, X.; Gautam, V.; Maras, A.; Poynton, E. F.; Tavangar, P.; Yang, V.; Van Santen, J. A.; Ghosh, R.; Sarma, S.; Knutson, E.; Sullivan, V.; Jystad, A. M.; Renslow, R.; Sumner, L. W.; Linington, R. G.; Cort, J. R. NP-MRD: The Natural Products Magnetic Resonance Database. *Nucleic Acids Res.* **2022**, *50* (D1), D665–D677. DOI: 10.1093/nar/gkab1052.

61. Egan, J. M.; van Santen, J. A.; Liu, D. Y.; Linington, R. G. Development of an NMR-Based Platform for the Direct Structural Annotation of Complex Natural Products Mixtures. *J. Nat. Prod.* **2021**, *84* (4), 1044–1055. DOI: 10.1021/ACS.JNATPROD.0C01076.

62. Howarth, A.; Goodman, J. M. The DP5 Probability, Quantification and Visualisation of Structural Uncertainty in Single Molecules. *Chem. Sci.* **2022**, *13* (12), 3507. DOI: 10.1039/D1SC04406K.

63. Kurita, K. L.; Glassey, E.; Linington, R. G. Integration of High-Content Screening and Untargeted Metabolomics for Comprehensive Functional Annotation of Natural Product Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (39), 11999–12004. DOI: 10.1073/pnas.1507743112.

64. Kurita, K. L.; Linington, R. G. Connecting Phenotype and Chemotype: High-Content Discovery Strategies for Natural Products Research. *J. Nat. Prod.* **2015**, *78* (3), 587–596. DOI: 10.1021/acs.jnatprod.5b00017.

65. Nothias, L. F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; Aicheler, F.; Aksenov, A. A.; Alka, O.; Allard, P. M.; Barsch, A.; Cachet, X.; Caraballo-Rodriguez, A. M.; Da Silva, R. R.; Dang, T.; Garg, N.; Gauglitz, J. M.; Gurevich, A.; Isaac, G.; Jarmusch, A. K.; Kameník, Z.; Kang, K. Bin; Kessler, N.; Koester, I.; Korf, A.; Le Gouellec, A.; Ludwig, M.; Martin H, C.; McCall, L. I.; McSayles, J.; Meyer, S. W.; Mohimani, H.; Morsy, M.; Moyne, O.; Neumann, S.; Neuweger, H.; Nguyen, N. H.; Nothias-Esposito, M.; Paolini, J.; Phelan, V. V.; Pluskal, T.; Quinn, R. A.; Rogers, S.; Shrestha, B.; Tripathi, A.; van der Hooft, J. J. J.; Vargas, F.; Weldon, K. C.; Witting, M.; Yang, H.; Zhang, Z.; Zubeil, F.; Kohlbacher, O.; Böcker, S.; Alexandrov, T.; Bandeira, N.; Wang, M.; Dorrestein, P. C. Feature-Based Molecular Networking in the GNPS Analysis Environment. *Nat. Methods.* **2020**, *17* (9), 905–908. DOI: 10.1038/S41592-020-0933-6.

66. McCaughey, C. S.; van Santen, J. A.; van der Hooft, J. J. J.; Medema, M. H.; Linington, R. G. An Isotopic Labeling Approach Linking Natural Products with Biosynthetic Gene Clusters. *Nat. Chem. Biol.* **2022**, *18* (3), 295–304. DOI: 10.1038/s41589-021-00949-6.

67. Blin, K.; Shaw, S.; Kloosterman, A. M.; Charlop-Powers, Z.; Van Wezel, G. P.; Medema, M. H.; Weber, T. AntiSMASH 6.0: Improving Cluster Detection and Comparison Capabilities. *Nucleic Acids Res.* **2021**, *49* (W1), W29. DOI: 10.1093/NAR/GKAB335.

68. de Jonge, N. F.; Louwen, J. J. R.; Chekmeneva, E.; Camuzeaux, S.; Vermeir, F. J.; Jansen, R. S.; Huber, F.; van der Hooft, J. J. J. MS2Query: Reliable and Scalable MS2 Mass Spectra-Based Analogue Search. *Nat. Commun.* **2023**, *14* (1), 1–12. DOI: 10.1038/s41467-023-37446-4.

69. Guerrero-Egido, G.; Pintado, A.; Bretscher, K. M.; Arias-Giraldo, L. M.; Paulson, J. N.; Spaink, H. P.; Claessen, D.; Ramos, C.; Cazorla, F. M.; Medema, M. H.; Raaijmakers, J. M.; Carrión, V. J. BacLIFE: A User-Friendly Computational Workflow for Genome Analysis and Prediction of Lifestyle-Associated Genes in Bacteria. *Nat. Commun.* **2024**, *15* (1), 1–18. DOI: 10.1038/s41467-024-46302-y.

70. de Jonge, N. F.; Mildau, K.; Meijer, D.; Louwen, J. J. R.; Bueschl, C.; Huber, F.; van der Hooft, J. J. J. Good Practices and Recommendations for Using and Benchmarking Computational Metabolomics Metabolite Annotation Tools. *Metabolomics.* **2022**, *18* (12), 1–22. DOI: 10.1007/S11306-022-01963-Y.

71. Li, C.; Chu, S.; Tan, S.; Yin, X.; Jiang, Y.; Dai, X.; Gong, X.; Fang, X.; Tian, D. Towards Higher Sensitivity of Mass Spectrometry: A Perspective From the Mass Analyzers. *Front. Chem.* **2021**, *9*, 813359. DOI: 10.3389/fchem.2021.813359.

72. Griffiths, J. A Brief History of Mass Spectrometry. *Anal. Chem.* **2008**, *80* (15), 5678–5683. DOI: 10.1021/ac8013065.

73. Ren, J.-L.; Zhang, A.-H.; Kong, L.; Wang, X.-J. Advances in Mass Spectrometry-Based Metabolomics for Investigation of Metabolites. *RSC Adv.* **2018**, *8* (40), 22335–22350. DOI: 10.1039/C8RA01574K.

74. Zhang, W.; Xu, L.; Zhang, H. Recent Advances in Mass Spectrometry Techniques for Atmospheric Chemistry Research on Molecular-level. *Mass Spectrom. Rev.* **2023**, 1–44. DOI: 10.1002/mas.21857.

75. Zhu, X.; Xu, T.; Peng, C.; Wu, S. Advances in MALDI Mass Spectrometry Imaging Single Cell and Tissues. *Front. Chem.* **2022**, *9*, 782432. DOI: 10.3389/fchem.2021.782432.

76. Liu, L.; Wang, Z.; Zhang, Q.; Mei, Y.; Li, L.; Liu, H.; Wang, Z.; Yang, L. Ion Mobility Mass Spectrometry for the Separation and Characterization of Small Molecules. *Anal. Chem.* **2023**, *95* (1), 134–151. DOI: 10.1021/ACS.ANALCHEM.2C02866.

77. Cooper-Shepherd, D. A.; Wildgoose, J.; Kozlov, B.; Johnson, W. J.; Tyldesley-Worster, R.; Palmer, M. E.; Hoyes, J. B.; McCullagh, M.; Jones, E.; Tonge, R.; Marsden-Edwards, E.; Nixon, P.; Verenchikov, A.; Langridge, J. I. Novel Hybrid Quadrupole-Multireflecting Time-of-Flight Mass Spectrometry System. *J. Am. Soc. Mass Spectrom.* **2023**, *34* (2), 264–272. DOI: 10.1021/jasms.2c00281.

78. Verdun, C. M.; Fuchs, T.; Harar, P.; Elbrächter, D.; Fischer, D. S.; Berner, J.; Grohs, P.; Theis, F. J.; Krahmer, F. Group Testing for SARS-CoV-2 Allows for Up to 10-Fold Efficiency Increase Across Realistic Scenarios and Testing Strategies. *Front. Public Health.* **2021**, *9*, 583377. DOI: 10.3389/fpubh.2021.583377.

79. Shental, N.; Levy, S.; Wuvshet, V.; Skorniakov, S.; Shalem, B.; Ottolenghi, A.; Greenshpan, Y.; Steinberg, R.; Edri, A.; Gillis, R.; Goldhirsh, M.; Moscovici, K.; Sachren, S.; Friedman, L. M.; Nesher, L.; Shemer-Avni, Y.; Porgador, A.; Hertz, T. Efficient High-Throughput SARS-CoV-2 Testing to Detect Asymptomatic Carriers. *Sci. Adv.* **2020**, *6* (37), 5961–5972. DOI: 10.1126/sciadv.abc5961.

80. Dorfman, R. The Detection of Defective Members of Large Populations. *Annals of Mathematical Statistics.* **1943**, *14* (4), 436–440. DOI: 10.1214/aoms/1177731363.

81. Du, D.-Z.; Hwang, F. K. Combinatorial Group Testing and Its Applications (Applied Mathematics). In Series on Applied Mathematics; World Scientific Publishing Company, 1999; Vol. 12, pp 1 – 336. DOI: 10.1142/4252.

82. Elkin, L. L.; Harden, D. G.; Saldanha, S.; Ferguson, H.; Cheney, D. L.; Pieniazek, S. N.; Maloney, D. P.; Zewinski, J.; O'Connell, J.; Banks, M. Just-in-Time Compound Pooling Increases Primary Screening Capacity without Compromising Screening Quality. *J. Biomol. Screen.* **2015**, *20* (5), 577–587. DOI: 10.1177/1087057115572988.

83. Kainkaryam, R. M.; Woolf, P. J. PoolHiTS: A Shifted Transversal Design Based Pooling Strategy for High-Throughput Drug Screening. *BMC Bioinformatics.* **2008**, *9*, 256. DOI: 10.1186/1471-2105-9-256.

84. Kainkaryam, R. M.; Woolf, P. J. Pooling in High-Throughput Drug Screening. *Curr. Opin. Drug. Discov. Devel.* **2009**, *12* (3), 339–350. PMCID: PMC3204799.

85. Kainkaryam, R. M.; Bruex, A.; Gilbert, A. C.; Schiefelbein, J.; Woolf, P. J. PoolMC: Smart Pooling of MRNA Samples in Microarray Experiments. *BMC Bioinformatics.* **2010**, *11*, 299–307. DOI: 10.1186/1471-2105-11-299.

86. Ohnesorge, N.; Sasore, T.; Hillary, D.; Alvarez, Y.; Carey, M.; Kennedy, B. N. Orthogonal Drug Pooling Enhances Phenotype-Based Discovery of Ocular Antiangiogenic Drugs in Zebrafish Larvae. *Front. Pharmacol.* **2019**, *10*, 508. DOI: 10.3389/fphar.2019.00508.

87. Devlin, J. J.; Liang, A.; Trinh, L.; Polokoff, M. A.; Senator, D.; Zheng, W.; Kondracki, J.; Kretschmer, P. J.; Morser, J.; Lipson, S. E.; Spann, R.; Loughlin, J. A.; Dunn, K. V.; Morrissey, M. M. High Capacity Screening of Pooled Compounds: Identification of the Active Compound Without Re-Assay of Pool Members. *Drug Dev. Res.* **1996**, *37* (2), 80–85. DOI: 10.1002/(SICI)1098-2299(199602)37:2<80::AID-DDR3>3.0.CO;2-H.

88. Thierry-Mieg, N. A New Pooling Strategy for High-Throughput Screening: The Shifted Transversal Design. *BMC Bioinformatics.* **2006**, *7*, 28. DOI: 10.1186/1471-2105-7-28.

89. Stancliffe, E.; Schwaiger-Haber, M.; Sindelar, M.; Murphy, M. J.; Soerensen, M.; Patti, G. J. An Untargeted Metabolomics Workflow That Scales to Thousands of Samples for Population-Based Studies. *Anal. Chem.* **2022**, *94* (50), 17370–17378. DOI: 10.1021/acs.analchem.2c01270.

90. Broeckling, C. D.; Beger, R. D.; Cheng, L. L.; Cumeras, R.; Cuthbertson, D. J.; Dasari, S.; Davis, W. C.; Dunn, W. B.; Evans, A. M.; Fernández-Ochoa, A.; Gika, H.; Goodacre, R.; Goodman, K. D.; Gouveia, G. J.; Hsu, P.-C.; Kirwan, J. A.; Kodra, D.; Kuligowski, J.; Lan, R. S.-L.; Monge, M. E.; Moussa, L. W.; Nair, S. G.; Reisdorph, N.; Sherrod, S. D.; Ulmer Holland, C.; Vuckovic, D.; Yu, L.-R.; Zhang, B.; Theodoridis, G.; Mosley, J. D. Current Practices in LC-MS Untargeted Metabolomics: A Scoping Review on the Use of Pooled Quality Control Samples. *Anal. Chem.* **2023**, *95* (51), 18645–18654. DOI: 10.1021/acs.analchem.3c02924.

91. Motlekar, N.; Diamond, S. L.; Napper, A. D. Evaluation of an Orthogonal Pooling Strategy for Rapid High-Throughput Screening of Proteases. *Assay Drug Dev. Technol.* **2008**, *6* (3), 395–405. DOI: 10.1089/adt.2007.110.

92. Leão, T.; Wang, M.; Moss, N.; da Silva, R.; Sanders, J.; Nurk, S.; Gurevich, A.; Humphrey, G.; Reher, R.; Zhu, Q.; Belda-Ferre, P.; Glukhov, E.; Whitner, S.; Alexander, K. L.; Rex, R.; Pevzner, P.; Dorrestein, P. C.; Knight, R.; Bandeira, N.; Gerwick, W. H.; Gerwick, L. A Multi-Omics Characterization of the Natural Product Potential of Tropical Filamentous Marine Cyanobacteria. *Mar. Drugs.* **2021**, *19* (1), 20. DOI: 10.3390/md19010020.

93. Yu, Q.; Xiao, H.; Jedrychowski, M. P.; Schweppe, D. K.; Navarrete-Perea, J.; Knott, J.; Rogers, J.; Chouchani, E. T.; Gygi, S. P. Sample Multiplexing for Targeted Pathway Proteomics in Aging Mice. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (18), 9723–9732. DOI: 10.1073/pnas.1919410117.

94. Demarque, D. P.; Dusi, R. G.; de Sousa, F. D. M.; Grossi, S. M.; Silvério, M. R. S.; Lopes, N. P.; Espindola, L. S. Mass Spectrometry-Based Metabolomics Approach in the Isolation of Bioactive Natural Products. *Sci. Rep.* **2020**, *10* (1), 1–9. DOI: 10.1038/s41598-020-58046-y.

95. Sinclair, I.; Stearns, R.; Pringle, S.; Wingfield, J.; Datwani, S.; Hall, E.; Ghislain, L.; Majlof, L.; Bachman, M. Novel Acoustic Loading of a Mass Spectrometer: Toward next-Generation High-Throughput MS Screening. *J. Lab. Autom.* **2016**, *21* (1), 19–26. DOI: 10.1177/2211068215619124.

96. Deng, K.; Lan, X.; Fang, Q.; Li, M.; Xie, G.; Xie, L. Untargeted Metabolomics Reveals Alterations in the Primary Metabolites and Potential Pathways in the Vegetative Growth of *Morchella sextelata*. *Front. Mol. Biosci.* **2021**, *8*, 42. DOI: 10.3389/FMOLB.2021.632341/BIBTEX.

97. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. *BMC Bioinformatics.* **2010**, *11*, 395. DOI: 10.1186/1471-2105-11-395.

98. Schulze, C. J.; Bray, W. M.; Woerhmann, M. H.; Stuart, J.; Lokey, R. S.; Linington, R. G. “Function-First” Lead Discovery: Mode of Action Profiling of Natural Product Libraries Using Image-Based Screening. *Chem. Biol.* **2013**, *20* (2), 285–295. DOI: 10.1016/j.chembiol.2012.12.007.

99. Hoshino, S.; Okada, M.; Awakawa, T.; Asamizu, S.; Onaka, H.; Abe, I. Mycolic Acid Containing Bacterium Stimulates Tandem Cyclization of Polyene Macrolactam in a Lake Sediment Derived Rare Actinomycete. *Org. Lett.* **2017**, *19* (18), 4992–4995. DOI: 10.1021/acs.orglett.7b02508.

100. Skellam, E. J.; Stewart, A. K.; Strangman, W. K.; Wright, J. L. C. Identification of Micromonolactam, a New Polyene Macrocyclic Lactam from Two Marine *Micromonospora* Strains Using Chemical and Molecular Methods: Clarification of the Biosynthetic Pathway from a Glutamate Starter Unit. *J. Antibiot.* **2013**, *66* (7), 431–441. DOI: 10.1038/ja.2013.34.

101. Wong, W. R.; Oliver, A. G.; Linington, R. G. Development of Antibiotic Activity Profile Screening for the Classification and Discovery of Natural Product Antibiotics. *Chem. Biol.* **2012**, *19* (11), 1483–1495. DOI: 10.1016/j.chembiol.2012.09.014.

102. D'Atri, V.; Fekete, S.; Clarke, A.; Veuthey, J. L.; Guillarme, D. Recent Advances in Chromatography for Pharmaceutical Analysis. *Anal. Chem.* **2019**, *91* (1), 210–239. DOI: 10.1021/acs.analchem.8b05026.

103. Clark, C. M.; Nguyen, L.; Pham, V. C.; Sanchez, L. M.; Murphy, B. T. Automated Microbial Library Generation Using the Bioinformatics Platform IDBac. *Molecules.* **2022**, *27* (7), 2038. DOI: 10.3390/molecules27072038.

104. Kalkreuter, E.; Pan, G.; Cepeda, A. J.; Shen, B. Targeting Bacterial Genomes for Natural Product Discovery. *Trends. Pharmacol. Sci.* **2020**, *41* (1), 13–26. DOI: 10.1016/j.tips.2019.11.002.

105. de Medeiros, L. S.; de Araújo Júnior, M. B.; Peres, E. G.; da Silva, J. C. I.; Bassicheto, M. C.; Di Gioia, G.; Veiga, T. A. M.; Koolen, H. H. F. Discovering New Natural Products Using Metabolomics-Based Approaches. In *Advances in Experimental Medicine and Biology*; Springer, Cham, 2023; Vol. 1439, pp 185–224. DOI: 10.1007/978-3-031-41741-2_8.

106. Zhou, X.; Zhang, W.; Ouyang, Z. Recent Advances in On-Site Mass Spectrometry Analysis for Clinical Applications. *Trends Anal. Chem.* **2022**, *149*, 116548. DOI: 10.1016/j.trac.2022.116548.

107. Murray, K. K. Resolution and Resolving Power in Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2022**, *33* (12), 2342–2347. DOI: 10.1021/JASMS.2C00216.

108. Géhin, C.; Holman, S. W. Advances in High-Resolution Mass Spectrometry Applied to Pharmaceuticals in 2020: A Whole New Age of Information. *Anal. Sci. Adv.* **2021**, *2* (3–4), 142–156. DOI: 10.1002/ANSA.202000149.

109. Recchia, M. J. J.; Baumeister, T. U. H.; Liu, D. Y.; Linington, R. G. MultiplexMS: A Mass Spectrometry-Based Multiplexing Strategy for Ultra-High-Throughput Analysis of Complex Mixtures. *Anal. Chem.* **2023**, *95* (32), 11908–11917. DOI: 10.1021/acs.analchem.3c00939.

110. Lee, H.-J.; Kremer, D. M.; Sajjakulnukit, P.; Zhang, L.; Lyssiotis, C. A. A Large-Scale Analysis of Targeted Metabolomics Data from Heterogeneous Biological Samples Provides Insights into Metabolite Dynamics. *Metabolomics.* **2019**, *15* (7), 103. DOI: 10.1007/s11306-019-1564-8.

111. Nagana Gowda, G. A.; Raftery, D. NMR-Based Metabolomics. In *Advances in Experimental Medicine and Biology*; Adv. Exp. Med. Biol., 2021; Vol. 1280, pp 19–37. DOI: 10.1007/978-3-030-51652-9_2.

112. Zhou, J.; Yin, Y. Strategies for Large-Scale Targeted Metabolomics Quantification by Liquid Chromatography-Mass Spectrometry. *Analyst.* **2016**, *141* (23), 6362–6373. DOI: 10.1039/C6AN01753C.

113. Alseekh, S.; Aharoni, A.; Brotman, Y.; Contrepois, K.; D'Auria, J.; Ewald, J.; C. Ewald, J.; Fraser, P. D.; Giavalisco, P.; Hall, R. D.; Heinemann, M.; Link, H.; Luo, J.; Neumann, S.; Nielsen, J.; Perez de Souza, L.; Saito, K.; Sauer, U.; Schroeder, F. C.; Schuster, S.; Siuzdak, G.; Skirycz, A.; Sumner, L. W.; Snyder, M. P.; Tang, H.; Tohge, T.; Wang, Y.; Wen, W.; Wu, S.; Xu, G.; Zamboni, N.; Fernie, A. R. Mass Spectrometry-Based Metabolomics: A Guide for Annotation, Quantification and Best Reporting Practices. *Nat. Methods.* **2021**, *18* (7), 747–756. DOI: 10.1038/S41592-021-01197-1.

114. Vervoort, N.; Goossens, K.; Baeten, M.; Chen, Q. Recent Advances in Analytical Techniques for High Throughput Experimentation. *Anal. Sci. Adv.* **2021**, *2* (3–4), 109–127. DOI: 10.1002/ANSA.202000155.

115. Liigand, J.; Wang, T.; Kellogg, J.; Smedsgaard, J.; Cech, N.; Kruve, A. Quantification for Non-Targeted LC/MS Screening without Standard Substances. *Sci. Rep.* **2020**, *10* (1), 5808. DOI: 10.1038/s41598-020-62573-z.

116. Kapoore, R. V.; Vaidyanathan, S. Towards Quantitative Mass Spectrometry-Based Metabolomics in Microbial and Mammalian Systems. *Philos. Trans. A Math. Phys. Eng. Sci.* **2016**, *374* (2079), 20150363. DOI: 10.1098/rsta.2015.0363.

117. Khamis, M. M.; Klemm, N.; Adamko, D. J.; El-Aneed, A. Comparison of Accuracy and Precision between Multipoint Calibration, Single Point Calibration, and Relative Quantification for Targeted Metabolomic Analysis. *Anal. Bioanal. Chem.* **2018**, *410* (23), 5899–5913. DOI: 10.1007/s00216-018-1205-5.

118. Baumeister, T. U. H.; Aadland, E.; Linington, R. G.; Kvalheim, O. M. Multivariate Pattern Analysis: A Method and Software to Reveal, Quantify, and Visualize Predictive Association Patterns in Multicollinear Data. *BMC Bioinformatics.* **2024**, *25* (51). DOI: 10.1186/s12859-024-05660-6.

119. Visconti, G.; Boccard, J.; Feinberg, M.; Rudaz, S. From Fundamentals in Calibration to Modern Methodologies: A Tutorial for Small Molecules Quantification in Liquid Chromatography–Mass Spectrometry Bioanalysis. *Anal. Chim. Acta.* **2023**, *1240*, 340711. DOI: 10.1016/J.ACA.2022.340711.

120. Ashrafzadeh, S.; Golding, G. B.; Ilie, S.; Ilie, L. Scoring Alignments by Embedding Vector Similarity. *Brief. Bioinform.* **2024**, *25* (3), 178. DOI: 10.1093/BIB/BBAE178.

121. Tounta, V.; Liu, Y.; Cheyne, A.; Larrouy-Maumus, G. Metabolomics in Infectious Diseases and Drug Discovery. *Mol. Omics.* **2021**, *17* (3), 376–393. DOI: 10.1039/D1MO00017A.

122. Lindemann, C.; Thomanek, N.; Hundt, F.; Lerari, T.; Meyer, H. E.; Wolters, D.; Marcus, K. Strategies in Relative and Absolute Quantitative Mass Spectrometry Based Proteomics. *Biol. Chem.* **2017**, *398* (5–6), 687–699. DOI: 10.1515/hsz-2017-0104.

123. Cuevas-Delgado, P.; Dudzik, D.; Miguel, V.; Lamas, S.; Barbas, C. Data-Dependent Normalization Strategies for Untargeted Metabolomics—a Case Study. *Anal. Bioanal. Chem.* **2020**, *412* (24), 6391–6405. DOI: 10.1007/s00216-020-02594-9.

124. Lee, J. Y.; Styczynski, M. P. Leveraging the Mass Balances of Cellular Metabolism to Infer Absolute Concentrations from Relative Abundance Metabolomics Data. *bioRxiv.* **2021**. DOI: 10.1101/2021.05.13.444095.

125. Lee, J. Y.; Han, Y.; Styczynski, M. P. Towards Inferring Absolute Concentrations from Relative Abundance in Time-Course GC-MS Metabolomics Data. *Mol. Omics.* **2023**, *19* (2), 126–136. DOI: 10.1039/D2MO00168C.

126. Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature.* **2020**, *585* (7825), 357–362. DOI: 10.1038/s41586-020-2649-2.

127. The pandas development team. Pandas-Dev/Pandas: Pandas. **2020**. DOI: 10.5281/zenodo.3509134.

128. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830. DOI: 10.5555/1953048.2078195.

129. Morehouse, N. J.; Clark, T. N.; McMann, E. J.; van Santen, J. A.; Haeckl, F. P. J.; Gray, C. A.; Linington, R. G. Annotation of Natural Product Compound Families Using Molecular Networking Topology and Structural Similarity Fingerprinting. *Nat. Commun.* **2023**, *14*, 308. DOI: 10.1038/s41467-022-35734-z.

130. Bristow, A. W. T.; Webb, K. S. Intercomparison Study on Accurate Mass Measurement of Small Molecules in Mass Spectrometry. *J. Am. Soc. Mass. Spectrom.* **2003**, *14* (10), 1086–1098. DOI: 10.1016/S1044-0305(03)00403-3.

131. Marshall, A. G.; Blakney, G. T.; Chen, T.; Kaiser, N. K.; Mckenna, A. M.; Rodgers, R. P.; Ruddy, B. M.; Xian, F. Mass Resolution and Mass Accuracy: How Much Is Enough? *Mass Spectrom. (Tokyo)* **2013**, *2*, 9. DOI: 10.5702/massspectrometry.S0009.

132. Oliver, S. G.; Winson, M. K.; Kell, D. B.; Baganz, F. Systematic Functional Analysis of the Yeast Genome. *Trends Biotechnol.* **1998**, *16* (9), 373–378. DOI: 10.1016/S0167-7799(98)01214-1.

133. Rathahao-Paris, E.; Alves, S.; Junot, C.; Tabet, J.-C. High Resolution Mass Spectrometry for Structural Identification of Metabolites in Metabolomics. *Metabolomics.* **2016**, *12* (10). DOI: 10.1007/s11306-015-0882-8.

134. Cui, M.; Cheng, C.; Zhang, L. High-Throughput Proteomics: A Methodological Mini-Review. *Lab. Invest.* **2022**, *102*, 1170–1181. DOI: 10.1038/s41374-022-00830-7.

135. Rozanova, S.; Barkovits, K.; Nikolov, M.; Schmidt, C.; Urlaub, H.; Marcus, K. Quantitative Mass Spectrometry-Based Proteomics: An Overview. *Methods Mol. Biol.* **2021**, *2228*, 85–116. DOI: 10.1007/978-1-0716-1024-4_8.

136. Clasen, M. A.; Kurt, L. U.; Santos, M. D. M.; Lima, D. B.; Liu, F.; Gozzo, F. C.; Barbosa, V. C.; Carvalho, P. C. Increasing Confidence in Proteomic Spectral Deconvolution through Mass Defect. *Bioinformatics.* **2022**, *38* (22), 5119–5120. DOI: 10.1093/bioinformatics/btac638.

137. Jaitly, N.; Mayampurath, A.; Littlefield, K.; Adkins, J. N.; Anderson, G. A.; Smith, R. D. Decon2LS: An Open-Source Software Package for Automated Processing and Visualization of High Resolution Mass Spectrometry Data. *BMC Bioinformatics.* **2009**, *10* (87), 1–15. DOI: 10.1186/1471-2105-10-87.

138. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J. MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis. *Nucleic Acids Res.* **2018**, *46*, W486–W494. DOI: 10.1093/nar/gky310.

139. Zarrouk, E.; Lenski, M.; Bruno, C.; Thibert, V.; Contreras, P.; Privat, K.; Ameline, A.; Fabresse, N. High-Resolution Mass Spectrometry: Theoretical and Technological Aspects. *Toxicologie Analytique et Clinique.* **2022**, *34* (1), 3–18. DOI: 10.1016/J.TOXAC.2021.11.002.

140. Tanaka, M.; Shibue, T. High-Resolution Mass Spectrometry for Beginners: A Laboratory Experiment for Organic Chemistry Students. *J. Chem. Educ.* **2023**, *100* (12), 4734–4740. DOI: 10.1021/acs.jchemed.3c00330.

141. Arthur, K. L.; Turner, M. A.; Reynolds, J. C.; Creaser, C. S. Increasing Peak Capacity in Nontargeted Omics Applications by Combining Full Scan Field Asymmetric Waveform Ion Mobility Spectrometry with Liquid Chromatography–Mass Spectrometry. *Anal. Chem.* **2017**, *89* (6), 3452–3459. DOI: 10.1021/acs.analchem.6b04315.

142. Geue, N. Modern Electrospray Ionization Mass Spectrometry Techniques for the Characterization of Supramolecules and Coordination Compounds. *Anal. Chem.* **2024**, *96* (19), 7332–7341. DOI: 10.1021/acs.analchem.4c01028.

143. Annesley, T. M. Ion Suppression in Mass Spectrometry. *Clin. Chem.* **2003**, *49* (7), 1041–1044. DOI: 10.1373/49.7.1041.

144. Pidot, S. J.; Coyne, S.; Kloss, F.; Hertweck, C. Antibiotics from Neglected Bacterial Sources. *Int. J. Med. Microbiol.* **2014**, *304* (1), 14–22. DOI: 10.1016/J.IJMM.2013.08.011.

145. Alkatheri, A. H.; Yap, P. S. X.; Abushelaibi, A.; Lai, K. S.; Cheng, W. H.; Erin Lim, S. H. Microbial Genomics: Innovative Targets and Mechanisms. *Antibiotics.* **2023**, *12* (2). DOI: 10.3390/ANTIBIOTICS12020190.

146. Kunakom, S.; Eustáquio, A. S. *Burkholderia* as a Source of Natural Products. *J. Nat. Prod.* **2019**, *82* (7), 2018–2037. DOI: 10.1021/acs.jnatprod.8b01068.

147. Compant, S.; Nowak, J.; Coenye, T.; Clément, C.; Ait Barka, E. Diversity and Occurrence of *Burkholderia* Spp. in the Natural Environment. *FEMS Microbiol. Rev.* **2008**, *32* (4), 607–626. DOI: 10.1111/J.1574-6976.2008.00113.X.

148. Partida-Martinez, L. P.; Hertweck, C. Pathogenic Fungus Harbours Endosymbiotic Bacteria for Toxin Production. *Nature.* **2005**, *437* (7060), 884–888. DOI: 10.1038/NATURE03997.

149. Scherlach, K.; Brendel, N.; Ishida, K.; Dahse, H. M.; Hertweck, C. Photochemical Oxazole–Nitrile Conversion Downstream of Rhizoxin Biosynthesis and Its Impact on Antimitotic Activity. *Org. Biomol. Chem.* **2012**, *10* (30), 5756–5759. DOI: 10.1039/C2OB25250C.

150. Schmitt, I.; Partida-Martinez, L. P.; Winkler, R.; Voigt, K.; Einax, E.; Dölz, F.; Telle, S.; Wöstemeyer, J.; Hertweck, C. Evolution of Host Resistance in a Toxin-Producing Bacterial–Fungal Alliance. *ISME J.* **2008**, *2* (6), 632–641. DOI: 10.1038/ismej.2008.19.

151. Cordova-Kreylos, A. L.; Fernandez, L. E.; Koivunen, M.; Yang, A.; Flor-Weiler, L.; Marrone, P. G. Isolation and Characterization of *Burkholderia rinojensis* Sp. Nov., a Non-*Burkholderia* Cepacia Complex Soil Bacterium with Insecticidal and Miticidal Activities. *Appl. Environ. Microbiol.* **2013**, *79* (24), 7669–7678. DOI: 10.1128/AEM.02365-13.

152. Nickzad, A.; Lépine, F.; Déziel, E. Quorum Sensing Controls Swarming Motility of *Burkholderia glumae* through Regulation of Rhamnolipids. *PLoS One.* **2015**, *10* (6). DOI: 10.1371/JOURNAL.PONE.0128509.

153. Nyanasegran, P. K.; Nathan, S.; Firdaus-Raih, M.; Azlan, N.; Muhammad, N.; Ng, C. L. Biofilm Signaling, Composition and Regulation in *Burkholderia pseudomallei*. *J. Microbiol. Biotechnol.* **2023**, *33* (1), 15. DOI: 10.4014/JMB.2207.07032.

154. Butt, A. T.; Thomas, M. S. Iron Acquisition Mechanisms and Their Role in the Virulence of *Burkholderia* Species. *Front. Cell. Infect. Microbiol.* **2017**, *7,* 460. DOI: 10.3389/FCIMB.2017.00460.

155. Lewenza, S.; Conway, B.; Greenberg, E. P.; Sokol, P. A. Quorum Sensing in *Burkholderia cepacia*: Identification of the LuxRI Homologs CepRI. *J. Bacteriol.* **1999**, *181* (3), 748–756. DOI: 10.1128/JB.181.3.748-756.1999.

156. Eustáquio, A. S.; Janso, J. E.; Ratnayake, A. S.; O'donnell, C. J.; Koehn, F. E. Spliceostatin Hemiketal Biosynthesis in *Burkholderia* Spp. Is Catalyzed by an Iron/α-Ketoglutarate-Dependent Dioxygenase. *Proc. Natl. Acad. Sci. U. S. A*. **2014**, *111* (33), E3376–E3385. DOI: 10.1073/pnas.1408300111.

157. Hansanant, N.; Smith, L. Occidiofungin: Actin Binding as a Novel Mechanism of Action in an Antifungal Agent. *Antibiotics.* **2022**, *11* (9). DOI: 10.3390/ANTIBIOTICS11091143.

158. Deng, P.; Foxfire, A.; Xu, J.; Baird, S. M.; Jia, J.; Delgado, K. H.; Shin, R.; Smith, L.; Lu, S. E. The Siderophore Product Ornibactin Is Required for the Bactericidal Activity of *Burkholderia contaminans* MS14. *Appl. Environ. Microbiol.* **2017**, *83* (8). DOI: 10.1128/AEM.00051-17.

159. Franke, J.; Ishida, K.; Hertweck, C. Evolution of Siderophore Pathways in Human Pathogenic Bacteria. *J. Am. Chem. Soc.* **2014**, *136* (15), 5599–5602. DOI: 10.1021/JA501597W.

160. Ghosh, A. K.; Mishevich, J. L.; Jurica, M. S. Spliceostatins and Derivatives: Chemical Syntheses and Biological Properties of Potent Splicing Inhibitors. *J. Nat. Prod.* **2021**, *84* (5), 1681. DOI: 10.1021/ACS.JNATPROD.1C00100.

161. Medema, M. H.; Cimermancic, P.; Sali, A.; Takano, E.; Fischbach, M. A. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. *PLoS Comput. Biol.* **2014**, *10* (12), e1004016. DOI: 10.1371/JOURNAL.PCBI.1004016.

162. Scherlach, K.; Hertweck, C. Mining and Unearthing Hidden Biosynthetic Potential. *Nat. Commun.* **2021**, *12* (1), 1–12. DOI: 10.1038/s41467-021-24133-5.

163. Chen, H.; Sun, T.; Bai, X.; Yang, J.; Yan, F.; Yu, L.; Tu, Q.; Li, A.; Tang, Y.; Zhang, Y.; Bian, X.; Zhou, H. Genomics-Driven Activation of Silent Biosynthetic Gene Clusters in *Burkholderia gladioli* by Screening Recombineering System. *Molecules.* **2021**, *26* (3). DOI: 10.3390/MOLECULES26030700.

164. Chen, H.; Bai, X.; Sun, T.; Wang, X.; Zhang, Y.; Bian, X.; Zhou, H. The Genomic-Driven Discovery of Glutarimide-Containing Derivatives from *Burkholderia gladioli*. *Molecules.* **2023**, *28* (19). DOI: 10.3390/molecules28196937.

165. Tomm, H. A.; Ucciferri, L.; Ross, A. C. Advances in Microbial Culturing Conditions to Activate Silent Biosynthetic Gene Clusters for Novel Metabolite Production. *J. Ind. Microbiol. Biotechnol.* **2019**, *46* (9–10), 1381–1400. DOI: 10.1007/S10295-019-02198-Y.

166. Bauman, K. D.; Butler, K. S.; Moore, B. S.; Chekan, J. R. Genome Mining Methods to Discover Bioactive Natural Products. *Nat. Prod. Rep.* **2021**, *38* (11), 2100. DOI: 10.1039/D1NP00032B.

167. Chevrette, M. G.; Gavrilidou, A.; Mantri, S.; Selem-Mojica, N.; Ziemert, N.; Barona-Gómez, F. The Confluence of Big Data and Evolutionary Genome Mining for the Discovery of Natural Products. *Nat. Prod. Rep.* **2021**, *38* (11), 2024–2040. DOI: 10.1039/D1NP00013F.

168. Romanowski, S. B.; Lee, S.; Kunakom, S.; Paulo, B. S.; Recchia, M. J. J.; Liu, D. Y.; Cavanagh, H.; Linington, R. G.; Eustáquio, A. S. Identification of the Lipodepsipeptide Selethramide Encoded in a Giant Nonribosomal Peptide Synthetase from a *Burkholderia* Bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120* (42), e2304668120. DOI: 10.1073/PNAS.2304668120.

169. Gavriilidou, A.; Kautsar, S. A.; Zaburannyi, N.; Krug, D.; Müller, R.; Medema, M. H.; Ziemert, N. Compendium of Specialized Metabolite Biosynthetic Diversity Encoded in Bacterial Genomes. *Nat. Microbiol.* **2022**, *7* (5), 726–735. DOI: 10.1038/s41564-022-01110-2.

170. Bader, C. D.; Panter, F.; Müller, R. In Depth Natural Product Discovery - Myxobacterial Strains That Provided Multiple Secondary Metabolites. *Biotechnol. Adv.* **2020**, *39*, 107480. DOI: 10.1016/J.BIOTECHADV.2019.107480.

171. Challis, G. L.; Hopwood, D. A. Synergy and Contingency as Driving Forces for the Evolution of Multiple Secondary Metabolite Production by *Streptomyces* Species. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (24), 14555–14561. DOI: 10.1073/PNAS.1934677100.

172. Hegemann, J. D.; Birkelbach, J.; Walesch, S.; Müller, R. Current Developments in Antibiotic Discovery: Global Microbial Diversity as a Source for Evolutionary Optimized Anti-bacterials. *EMBO Rep.* **2023**, *24* (1), e56184. DOI: 10.15252/EMBR.202256184.

173. Adaikpoh, B. I.; Fernandez, H. N.; Eustáquio, A. S. Biotechnology Approaches for Natural Product Discovery, Engineering and Production Based on *Burkholderia* Bacteria. *Curr. Opin. Biotechnol.* **2022**, *77*, 102782. DOI: 10.1016/J.COPBIO.2022.102782.

174. Huo, L.; Hug, J. J.; Fu, C.; Bian, X.; Zhang, Y.; Müller, R. Heterologous Expression of Bacterial Natural Product Biosynthetic Pathways. *Nat. Prod. Rep.* **2019**, *36* (10), 1412–1436. DOI: 10.1039/C8NP00091C.

175. Wang, G.; Zhao, Z.; Ke, J.; Engel, Y.; Shi, Y. M.; Robinson, D.; Bingol, K.; Zhang, Z.; Bowen, B.; Louie, K.; Wang, B.; Evans, R.; Miyamoto, Y.; Cheng, K.; Kosina, S.; De Raad, M.; Silva, L.; Luhrs, A.; Lubbe, A.; Hoyt, D. W.; Francavilla, C.; Otani, H.; Deutsch, S.; Washton, N. M.; Rubin, E. M.; Mouncey, N. J.; Visel, A.; Northen, T.; Cheng, J. F.; Bode, H. B.; Yoshikuni, Y. CRAGE Enables Rapid Activation of Biosynthetic Gene Clusters in Undomesticated Bacteria. *Nat. Microbiol.* **2019**, *4* (12), 2498–2510. DOI: 10.1038/s41564-019-0573-8.

176. Fernandez, H. N.; Kretsch, A. M.; Kunakom, S.; Kadjo, A. E.; Mitchell, D. A.; Eustáquio, A. S. High-Yield Lasso Peptide Production in a *Burkholderia* Bacterial Host by Plasmid Copy Number Engineering. *ACS Synth. Biol.* **2024**, *13* (1), 337–350. DOI: 10.1021/ACSSYNBIO.3C00597.

177. Fergusson, C. H.; Coloma, J. M. F.; Valentine, M. C.; Haeckl, F. P. J.; Linington, R. G. Custom Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometric Database for Identification of Environmental Isolates of the Genus *Burkholderia* and Related Genera. *Appl. Environ. Microbiol.* **2020**, *86* (11). DOI: 10.1128/AEM.00354-20.

178. Haeckl, F. P. J.; Baldim, J. L.; Iskakova, D.; Kurita, K. L.; Soares, M. G.; Linington, R. G. A Selective Genome-Guided Method for Environmental *Burkholderia* Isolation. *J. Ind. Microbiol. Biotechnol.* **2019**, *46* (3–4), 345–362. DOI: 10.1007/S10295-018-02121-X.

179. Clark, C. M.; Costa, M. S.; Sanchez, L. M.; Murphy, B. T. Coupling MALDI-TOF Mass Spectrometry Protein and Specialized Metabolite Analyses to Rapidly Discriminate Bacterial Function. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (19), 4981–4986. DOI: 10.1073/PNAS.1801247115.

180. Zheng, W.; Wang, X.; Zhou, H.; Zhang, Y.; Li, A.; Bian, X. Establishment of Recombineering Genome Editing System in *Paraburkholderia megapolitana* Empowers Activation of Silent Biosynthetic Gene Clusters. *Microb. Biotechnol.* **2020**, *13* (2), 397. DOI: 10.1111/1751-7915.13535.

181. Marfey, P. Determination of D-Amino Acids. II. Use of a Bifunctional Reagent, 1,5-Difluoro-2,4-Dinitrobenzene. *Carlsberg Res. Commun.* **1984**, *49* (6), 591–596. DOI: 10.1007/BF02908688.

Appendix A.

Supplemental Figures and Tables

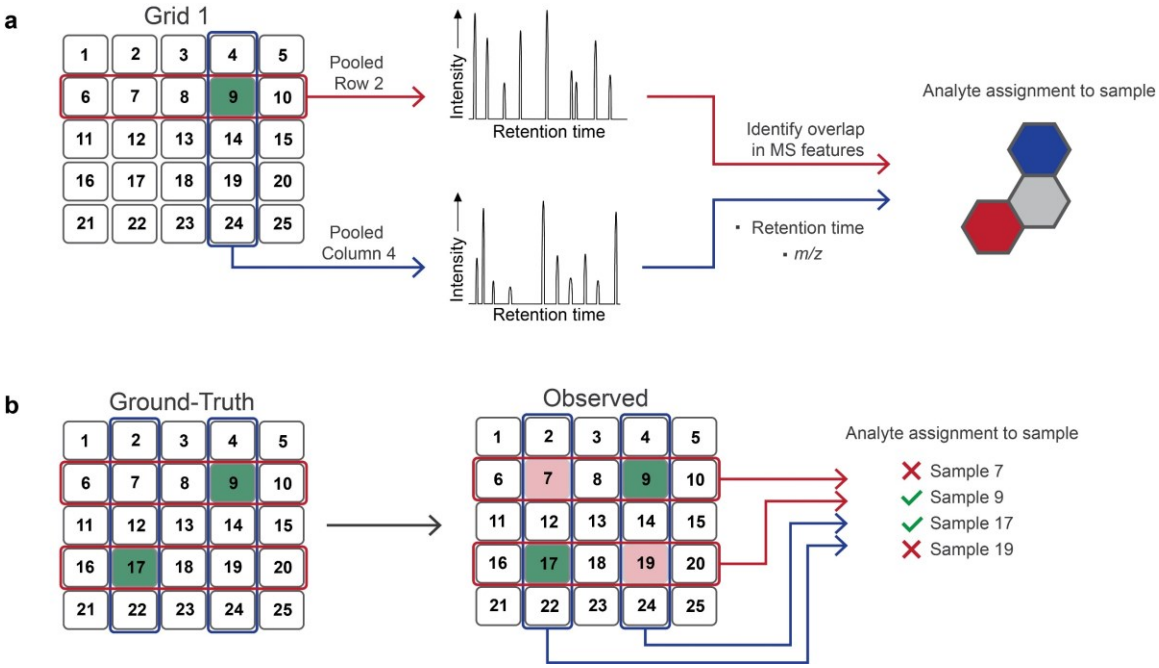


Figure A.1 **The Orthogonal Sampling Scheme Using One Grid.**
(**A**) MS feature deconvolution protocol using one grid. This illustrates a "successful solve" when a feature is only present in one well. (**B**) The depiction of false positive assignments when a feature is present in more than one well using a single sampling grid.

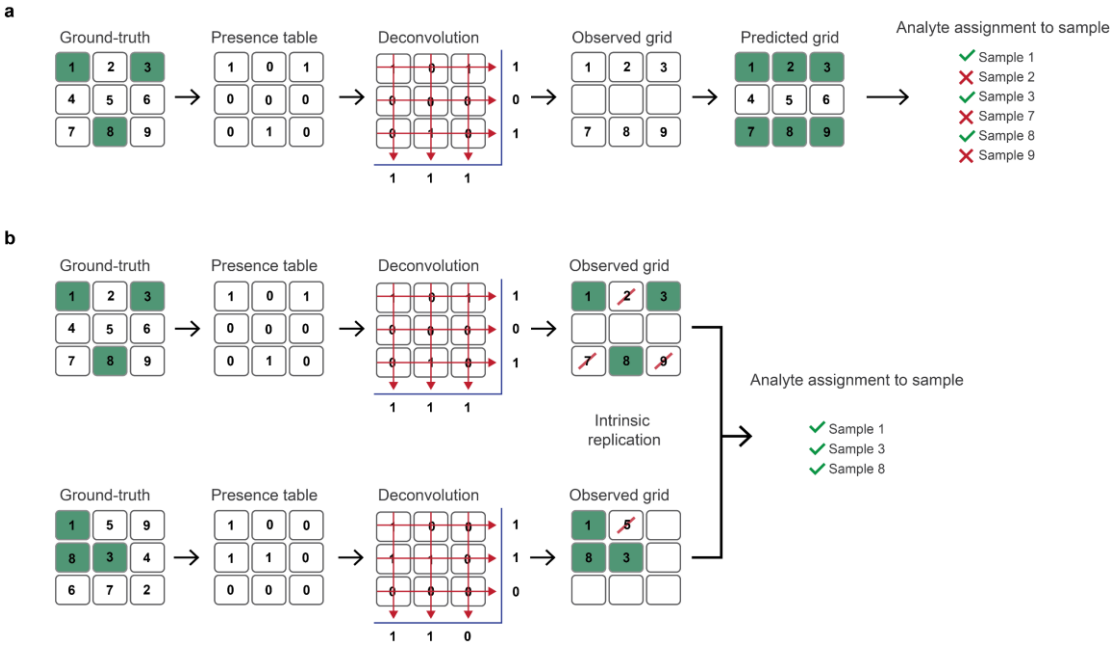


Figure A.2 ***In Silico* Computational Deconvolution Scheme (Single vs. Dual).** (**A**) The deconvoluted protocol using a single grid. A theoretical sampling grid is generated with a known analyte location. Under the deconvolution protocol, if an MS feature is present in a row or column, the entire row/column is given a 1. The whole row/column is given a 0 if no analyte is detected. When a row and column have a 1, then the sample at the intersection is given a 1. This process deconvolutes the sampling grid to create an observed grid. If there is nothing to compare to the observed grid, then this automatically becomes the predicted grid, creating an overinflated MS feature list with many false positives. These false positive assignments occur when $x > 1$. (**B**) MultiplexMS incorporates a dual-grid strategy to improve MS feature assignment. First, the same GT grid is rearranged using the MMSO protocol. Both grids are computationally deconvoluted to create observed grids for each arrangement. MS features in both observed grids are conserved, and the features present in only one grid are discarded from the final list of assigned features. This strategy provides a distinct advantage when $x > 1$.

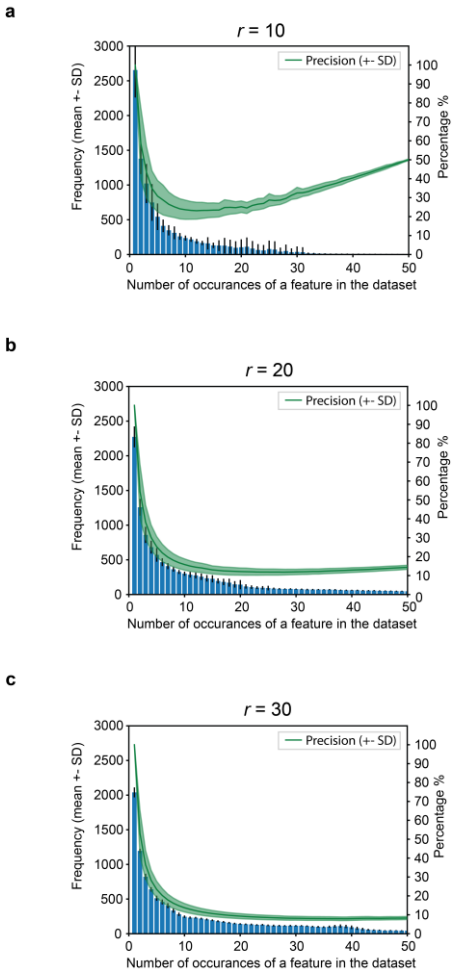


Figure A.3 **Single-Grid Orthogonal Sampling Disadvantage When *x* > 1.** (**A**) *r* = 10, single-grid deconvolution, feature frequency count, and precision calculation. The precision of the MS feature assignment drops immediately after *x* > 1. (**B**) *r* = 20 shows no improvement in precision after *x* >1. (**C**) *r* = 30 shows no improvement in precision after *x* > 1.

Grid #1

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

	Column 1	Column 2	Column 3	Column 4	Column 5
Row 1	Lincomycin	Doxycycline	Ceftazidime	Midecamycin	Florfenicol
Row 2	Nystatin	Amoxicillin	Ampicillin	Amphotericin B	Novobiocin
Row 3	Chloramphenicol	Azithromycin	Cefadroxil	Penicillin G	Deferoxamine
Row 4	Cyclosporine A	Cloxacillin	Cycloheximide	Tetracycline	Clarithromycin
Row 5	Erythromycin	Clindamycin	Roxithromycin	Piperacillin	Mupirocin

Grid #2

1	7	13	19	25
12	18	24	5	6
23	4	10	11	17
9	15	16	22	3
20	21	2	8	14

	Column 1	Column 2	Column 3	Column 4	Column 5
Row 1	Lincomycin	Amoxicillin	Cefadroxil	Tetracycline	Mupirocin
Row 2	Azithromycin	Cycloheximide	Piperacillin	Florfenicol	Nystatin
Row 3	Roxithromycin	Midecamycin	Novobiocin	Chloramphenicol	Cloxacillin
Row 4	Amphotericin B	Deferoxamine	Cyclosporine A	Clindamycin	Ceftazidime
Row 5	Clarithromycin	Erythromycin	Doxycycline	Ampicillin	Penicillin G

Figure A.4 “One-Compound-One-Well” 5 x 5 Multiplexing Experiment. The arrangement of samples in the 5 x 5 multiplexing experiment using 25 NP standards. The initial and rearranged grids were organized following the MMSO protocol.

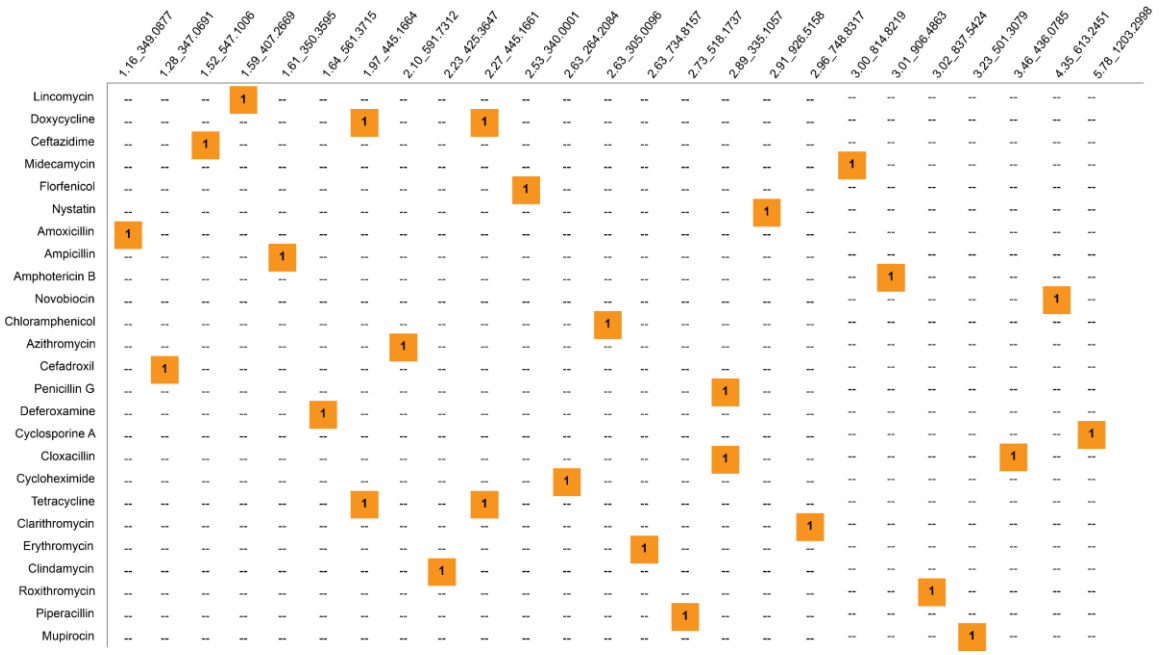


Figure A.5 MS Feature Presence Table Following Computational Deconvolution of Samples in the 5 x 5 Multiplexing Experiment.

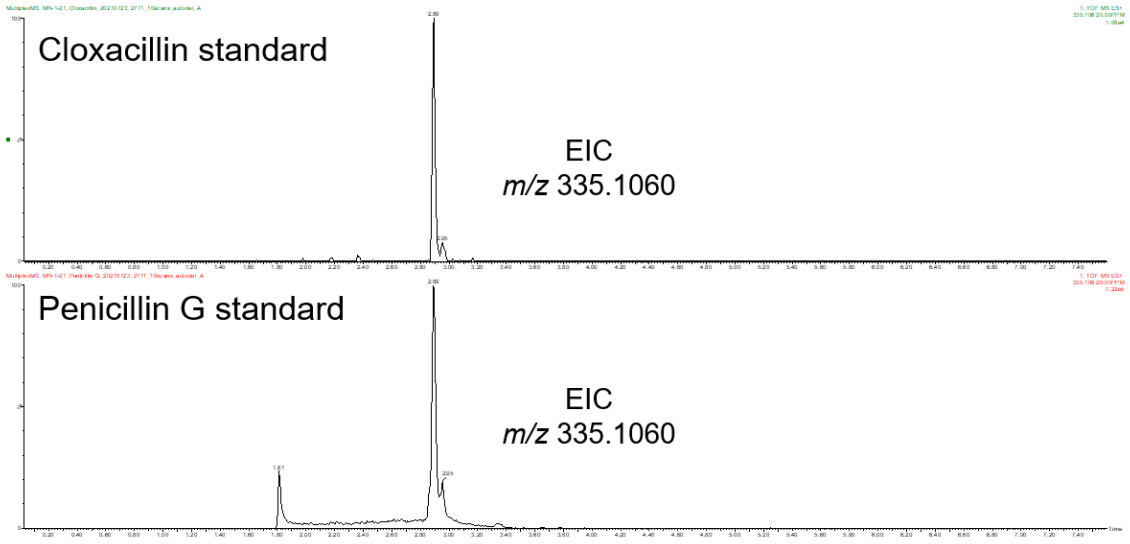


Figure A.6 EIC of *m/z* 335.1060 in the Penicillin G and Cloxacillin Standard. Computational deconvolution of the NP standards in the 5 x 5 experiment showed the presence of penicillin G in the cloxacillin standard. EIC of the appropriate mass shows contamination of penicillin G in the cloxacillin sample.

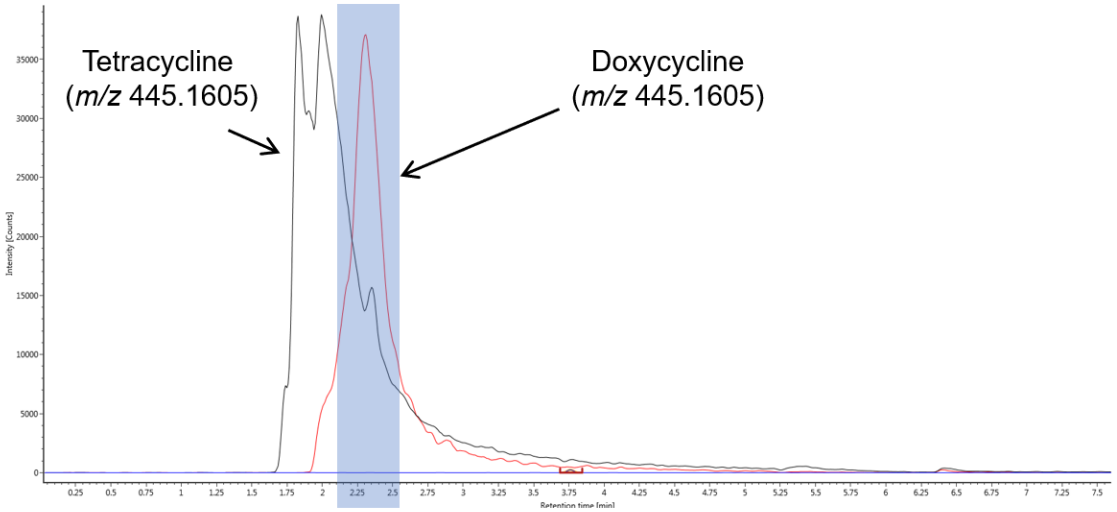


Figure A.7 **EIC of *m/z* 445.1605 in the Tetracycline and Doxycycline Standard.** Computational deconvolution of the NP standards in the 5 x 5 experiment showed the presence of tetracycline and doxycycline in each reconstructed sample. Tetracycline and doxycycline are structural isomers, and because of the close elution times and poor peak shape, the preprocessing software classified these molecules as the same. Therefore, each retention time_*m/z* pair was assigned to each reconstructed sample.

Grid 1

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
Row 1	Lincomycin	RLUS-1904C	RLUS2029D	RLUS-2024E	RLUS-1943F	RLUS-1929B	RLUS-1915C	RLUS-2039D	RLUS-2034E	RLCA-1165F
Row 2	RLUS-2028C	Deferoxamine	RLUS-2054E	RLUS-1916F	RLCA-1182B	RLUS-2052C	RLUS-1926D	RLUS-2011E	RLUS-2030F	RLUS-1933B
Row 3	RLCA-1161D	RLUS-2047E	Tetracycline	RLUS-2026B	RLUS-2048C	RLUS-1953D	RLUS-1946E	RLUS-1945F	RLUS-2032B	RLUS-2055C
Row 4	RLUS-2025E	RLCA-1189F	RLUS-1918B	Azithromycin	RLUS-2036D	RLUS-2043E	RLCA-1159F	RLUS-1937B	RLUS-2019C	RLUS-2027D
Row 5	RLUS-1935F	RLUS-2020B	RLUS-1928C	RLUS-2007D	Clindamycin	RLCA-1160F	RLUS-2040B	RLCA-1172C	RLUS-2021D	RLCA-1173E
Row 6	RLUS-1903B	RLCA-1192C	RLUS-2031D	RLUS-1936E	RLUS-1949F	Erythromycin	RLCA-1202C	RLCA-1169D	RLCA-1201E	RLUS-2038F
Row 7	RLUS-2044C	RLUS-1939D	RLUS-2037E	RLUS-2022F	RLUS-1938B	RLUS-1934C	Nystatin	RLCA-1181E	RLCA-1163F	RLCA-1198B
Row 8	RLUS-2041D	RLUS-1906E	RLCA-1195F	RLUS-2046B	RLCA-1157C	RLUS-2016D	RLUS-2049E	Roxithromycin	RLUS-2008B	RLUS-1954C
Row 9	RLUS-2047E	RLUS-2018F	RLUS-1951B	RLUS-2012C	RLCA-1176D	RLUS-2057E	RLUS-2009F	RLCA-1186B	Mupirocin	RLCA-1158D
Row 10	RLUS-2023F	RLUS-2053B	RLUS-1950C	RLUS-2010D	RLUS-2056E	RLUS-2017F	RLUS-2014B	RLUS-2024C	RLCA-1204D	Novobiocin

Grid 2

1	12	23	34	45	56	67	78	89	100
22	33	44	55	66	77	88	99	10	11
43	54	65	76	87	98	9	20	21	32
64	75	86	97	8	19	30	31	42	53
85	96	7	18	29	40	41	52	63	74
6	17	28	39	50	51	62	73	84	95
27	38	49	60	61	72	83	94	5	16
48	59	70	71	82	93	4	15	26	37
69	80	81	92	3	14	25	36	47	58
90	91	2	13	24	35	46	57	68	79

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
Row 1	Lincomycin	Deferoxamine	Tetracycline	Azithromycin	Clindamycin	Erythromycin	Nystatin	Roxithromycin	Mupirocin	Novobiocin
Row 2	RLUS-2047E	RLUS-1918B	RLUS-2007D	RLUS-1949F	RLUS-1934C	RLUS-2049E	RLCA-1186B	RLCA-1204D	RLCA-1165F	RLUS-2028C
Row 3	RLUS-1928C	RLUS-1936E	RLUS-1938B	RLUS-2016D	RLUS-2009F	RLUS-2024C	RLUS-2034E	RLUS-1933B	RLCA-1161D	RLCA-1189F
Row 4	RLUS-2022F	RLCA-1157C	RLUS-2057E	RLUS-2014B	RLUS-2039D	RLUS-2030F	RLUS-2055C	RLUS-2025E	RLUS-2020B	RLUS-2031D
Row 5	RLCA-1176D	RLUS-2017F	RLUS-1915C	RLUS-2011E	RLUS-2032B	RLUS-2027D	RLUS-1935F	RLCA-1192C	RLUS-2037E	RLUS-2046B
Row 6	RLUS-1929B	RLUS-1926D	RLUS-1945F	RLUS-2019C	RLCA-1173E	RLUS-1903B	RLUS-1939D	RLCA-1195F	RLUS-2012C	RLUS-2056E
Row 7	RLUS-1946E	RLUS-1937B	RLUS-2021D	RLUS-2038F	RLUS-2044C	RLUS-1906E	RLUS-1951B	RLUS-2010D	RLUS-1943F	RLUS-2052C
Row 8	RLCA-1172C	RLCA-1201E	RLCA-1198B	RLUS-2041D	RLUS-2018F	RLUS-1950C	RLUS-2042E	RLCA-1182B	RLUS-1953D	RLCA-1159F
Row 9	RLCA-1163F	RLUS-1954C	RLUS-1947E	RLUS-2053B	RLUS-2029D	RLUS-1916F	RLUS-2048C	RLUS-2043E	RLUS-2040B	RLCA-1169D
Row 10	RLCA-1158D	RLUS-2023F	RLUS-1904C	RLUS-2054E	RLUS-2026B	RLUS-2036D	RLCA-1160F	RLCA-1202C	RLCA-1181E	RLUS-2008B

Figure A.8 Initial and Rearranged Preparation Tables for the 10 x 10 Multiplexing Experiment.

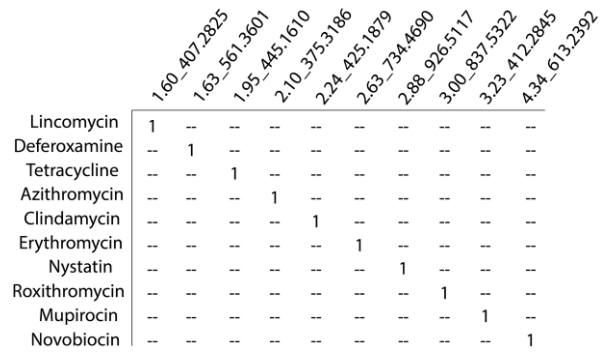


Figure A.9 MS Feature Deconvolution Performance for NPs in Complex Mixtures.

Results of the successful computational deconvolution of the ten standards used in the 10 x 10 multiplexing experiment. Each NP standard was correctly assigned with zero instances of false positive assignments.

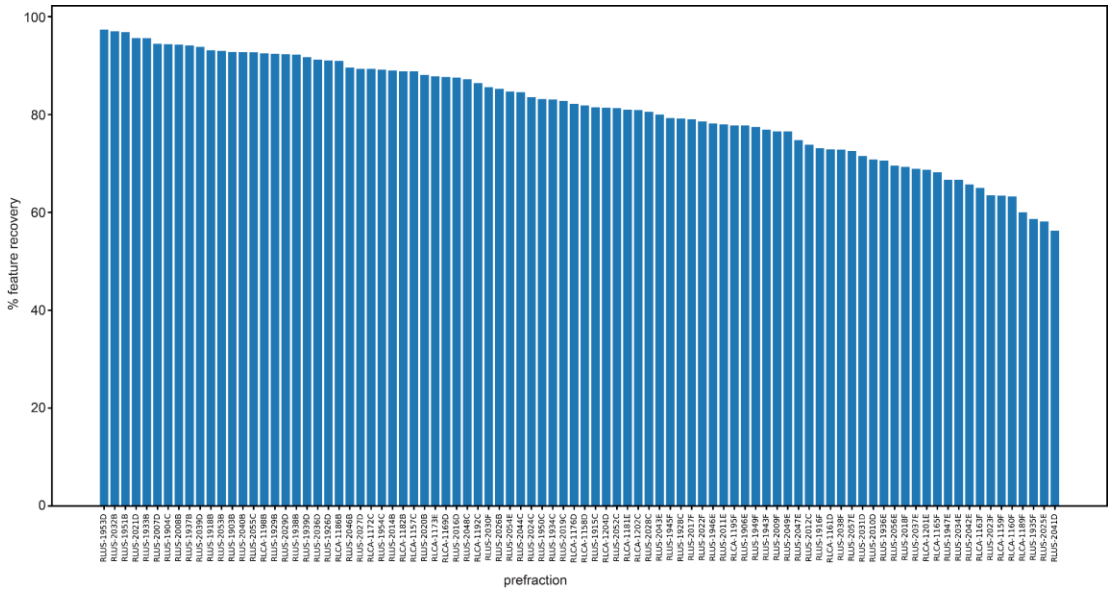


Figure A.10 Percent Information Recovery of the Reconstructed 90 Prefractions Versus the GT MS Feature List.

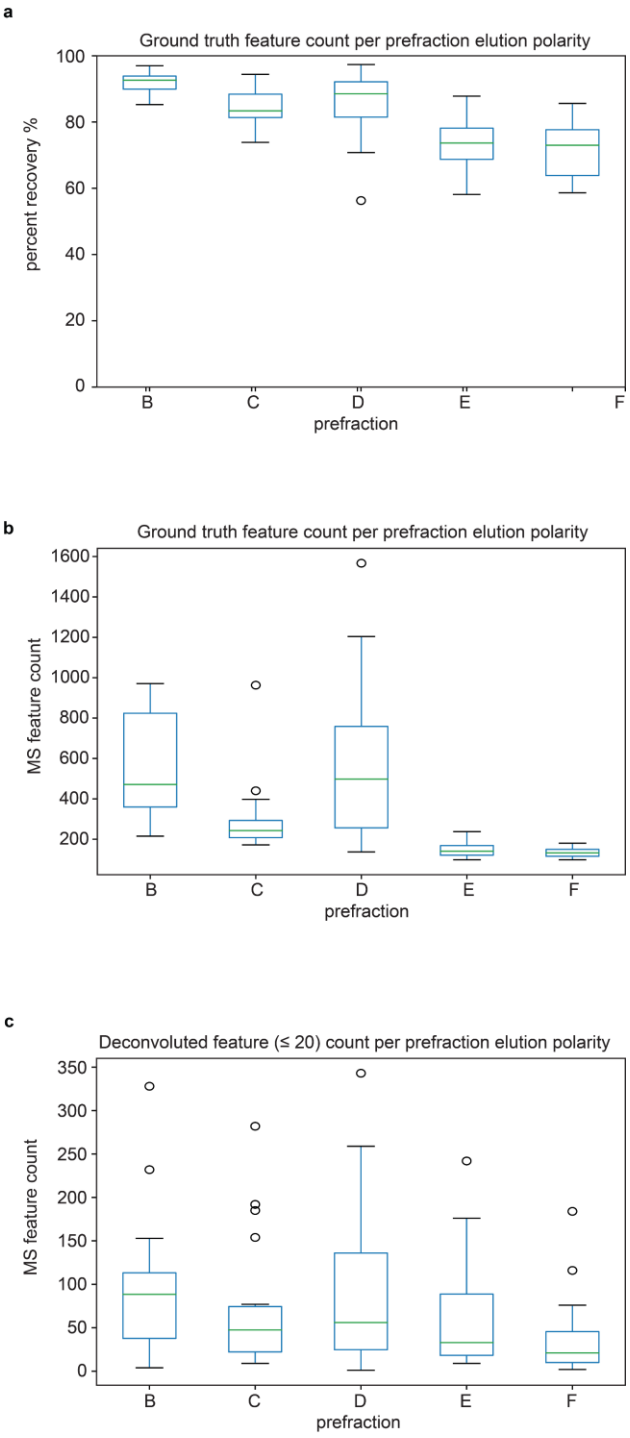


Figure A.11 Examination of a Feature Presence in GT Versus Reconstructed Samples for the 10 x 10 Multiplexing Experiment.
(**A**) The percent information recovery for MS feature comparison between the reconstructed and GT datasets per elution polarity of the prefractions. (**B**) An in-depth look at the MS feature count per prefraction elution polarity in the GT dataset. (**C**) MS feature counts in the reconstructed prefractions. Features in >20 samples were omitted from the final MS feature table.

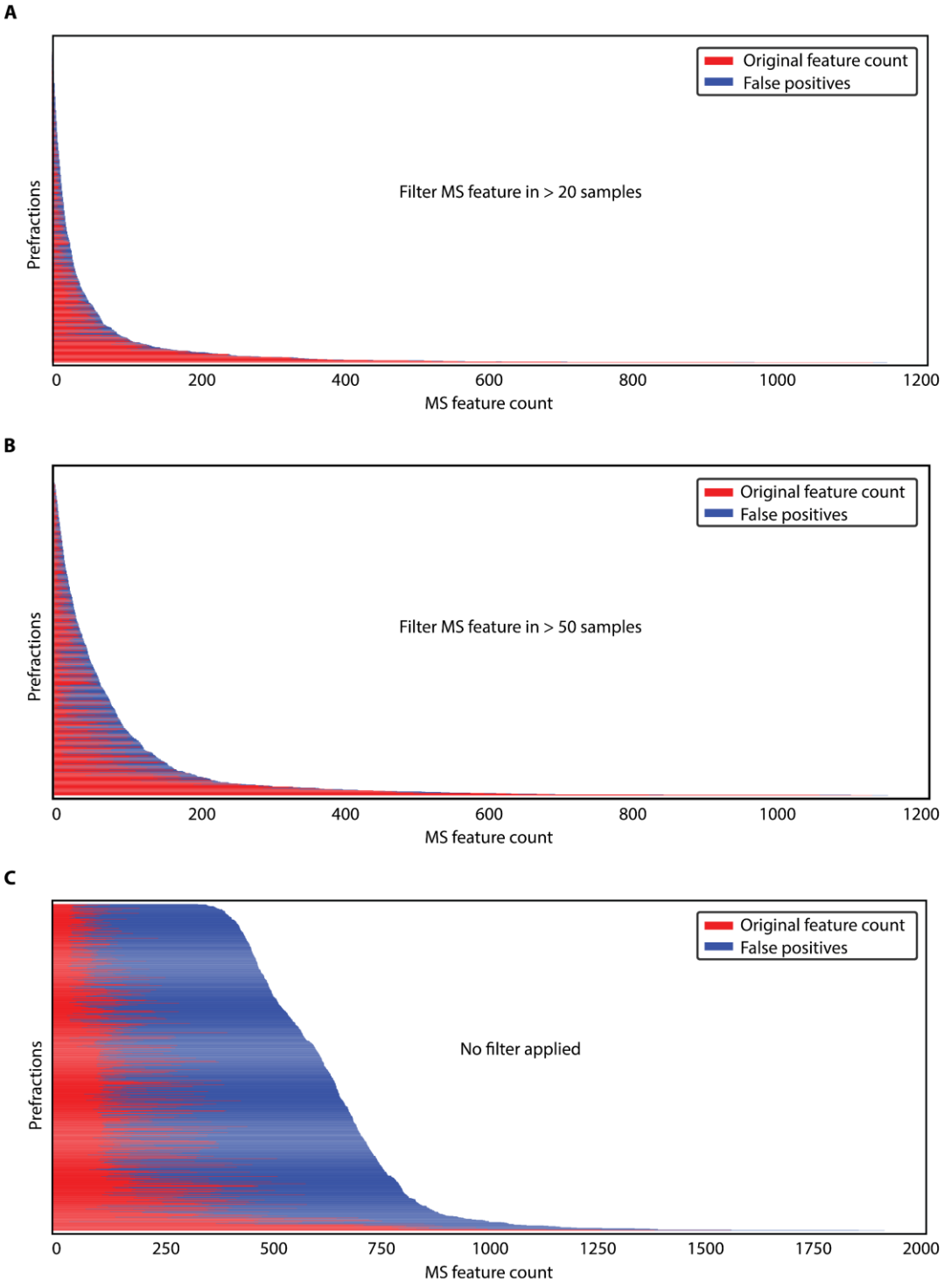


Figure A.12 Assessment of FPR Per Prefraction Following *In Silico* BioMAP Multiplexing Experiment.
Original feature count (red) per prefraction versus the reconstructed MS features (blue). (**A**) An MS feature count filter of >20 was applied, omitting those MS features that are present in more than 20 samples. (**B**) An MS feature count filter of >50 was applied, omitting those MS features that are present in more than 50 samples. (**C**) No filter is applied.

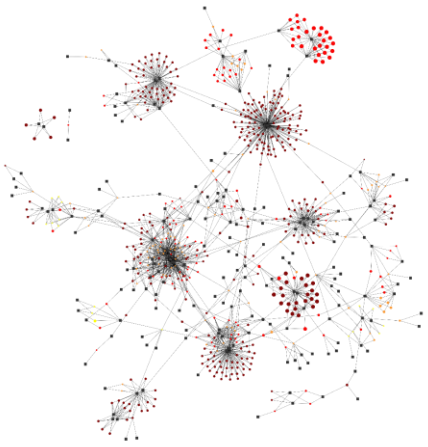


Figure A.13 High-Resolution Version of the NP Analyst Network from the Multiplexing Experiment (Supports Figure 2.8).

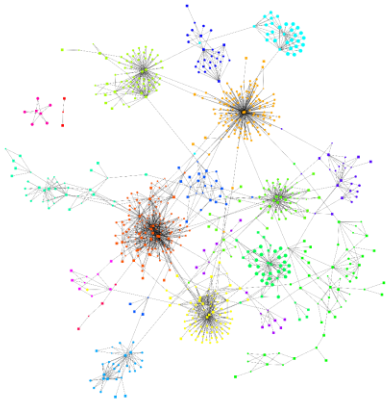


Figure A.14 High-Resolution Version of the NP Analyst Network from the Multiplexing Experiment (Colored by Community).

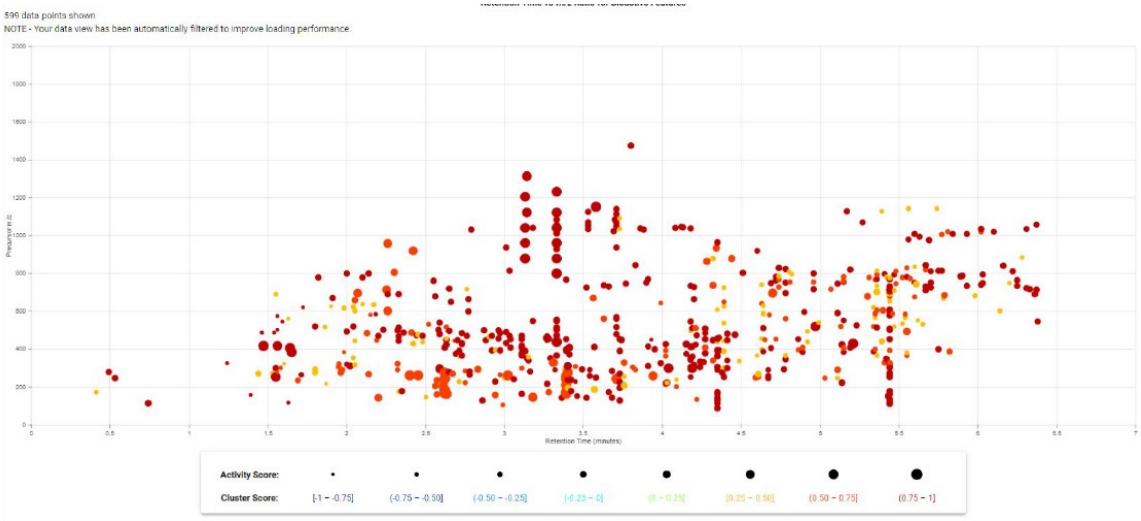


Figure A.15 The Scatter Plot Visualization of the NP Analyst Network from the Multiplexing Experiment.

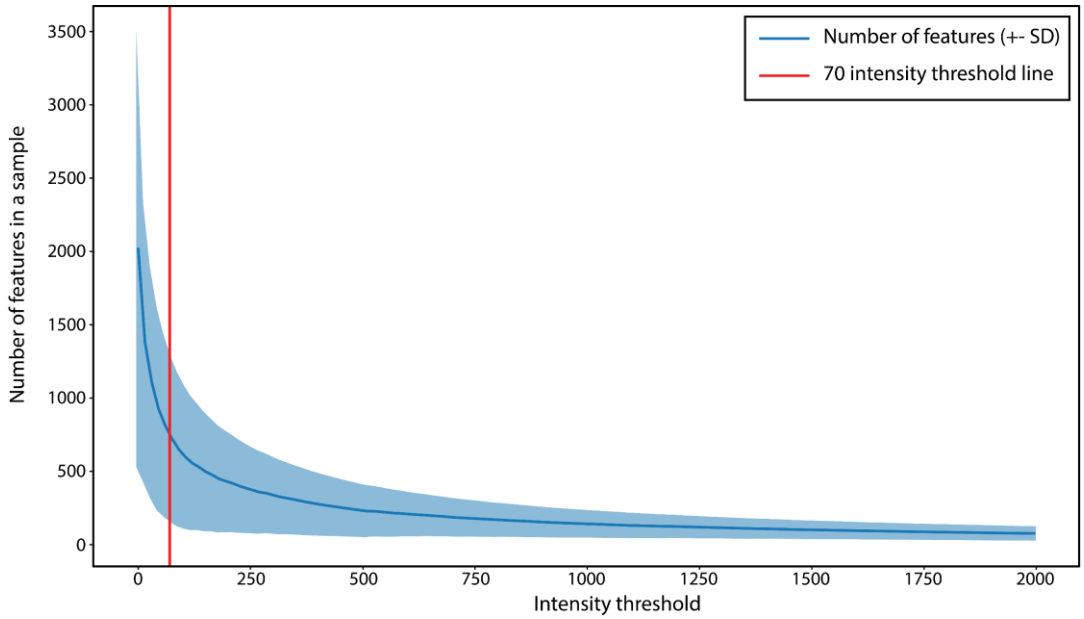


Figure A.16 Minimum Intensity Threshold Determination.
The minimum intensity cut-off threshold was implemented so that a feature would be deemed absent below a certain intensity. MS feature counts were determined on a per-sample basis versus the intensity of the detected features. The cut-off threshold was applied to the value indicated at the above inflection point. This greatly reduced ubiquitous “junk” features in the dataset.

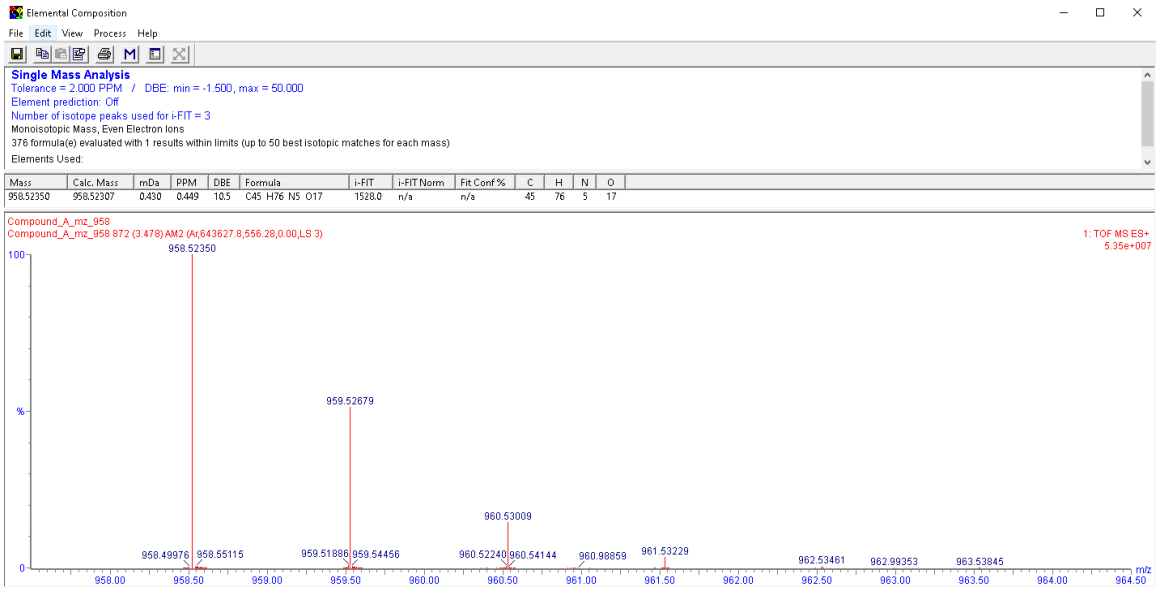


Figure A.17 HR-ESI-MS Data for Megapolipeptin A (5.2).

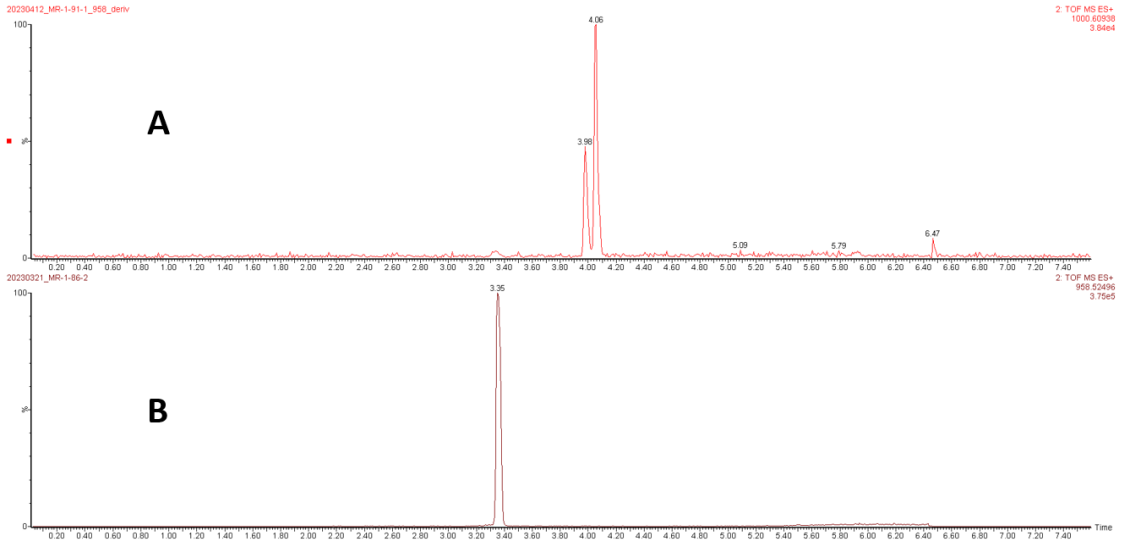


Figure A.18 Comparative Extracted Ion Chromatograms of Megapolipeptin A (5.2) and Methyl Ester Derivatives.
(**A**) EIC of the megapolipeptin A methyl ester derivatives (**5.4** and **5.5**). Two products were observed and isolated by HPLC. (**B**) EIC of purified megapolipeptin A (**5.2**).

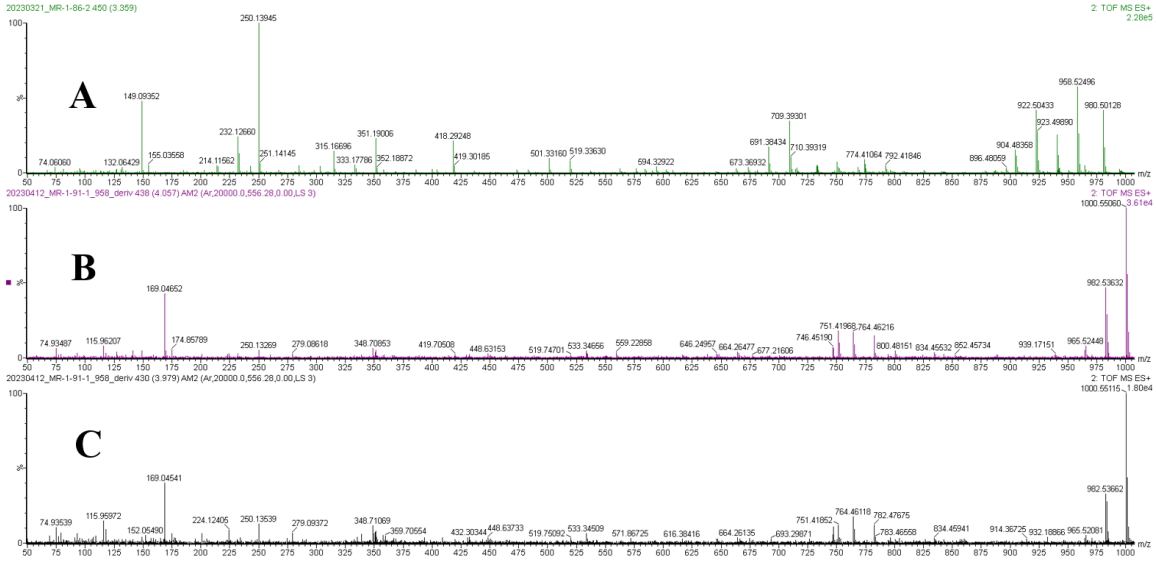


Figure A.19 Comparative MS Spectra of Megapolipeptin A (5.2) and Methyl Ester Derivatives.
(**A**) megapolipeptin A (**5.2**). (**B**) and (**C**) the two megapolipeptin A methyl ester derivatives (**B**: **5.4**, t_R 4.06 min; **C**: **5.5**, t_R 3.98 min).

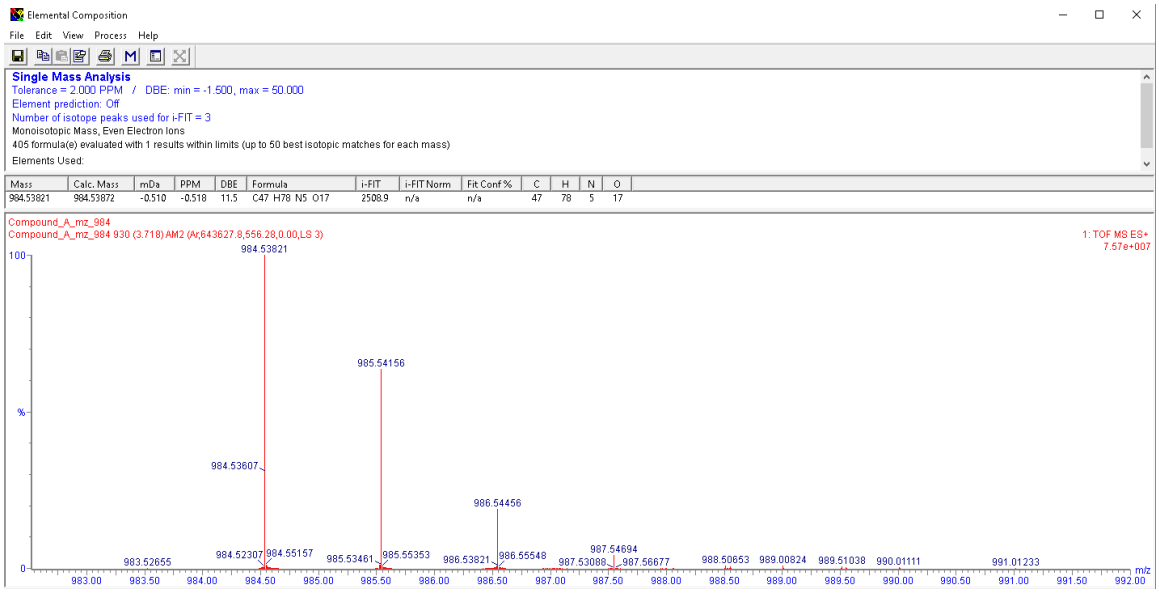


Figure A.20 HR-ESI-MS Data for Megapolipeptin B (5.3).

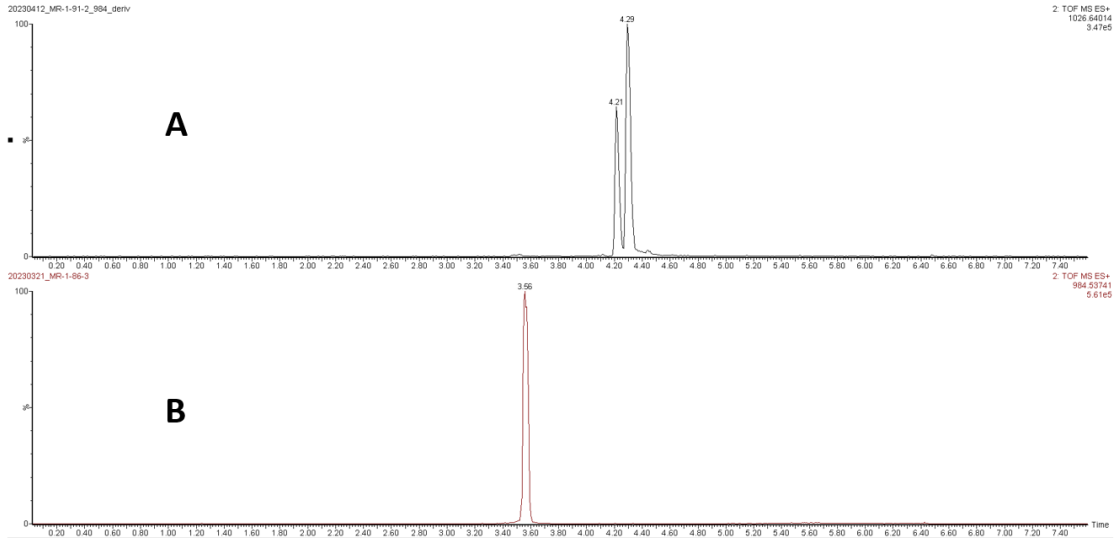


Figure A.21 Comparative Extracted Ion Chromatograms of Megapolipeptin B (5.3) and Methyl Ester Derivatives.
(**A**) EIC of the megapolipeptin B methyl ester derivatives (**5.6** and **5.7**). Two products were observed and isolated by HPLC. (**B**) EIC of purified megapolipeptin B (**5.3**).

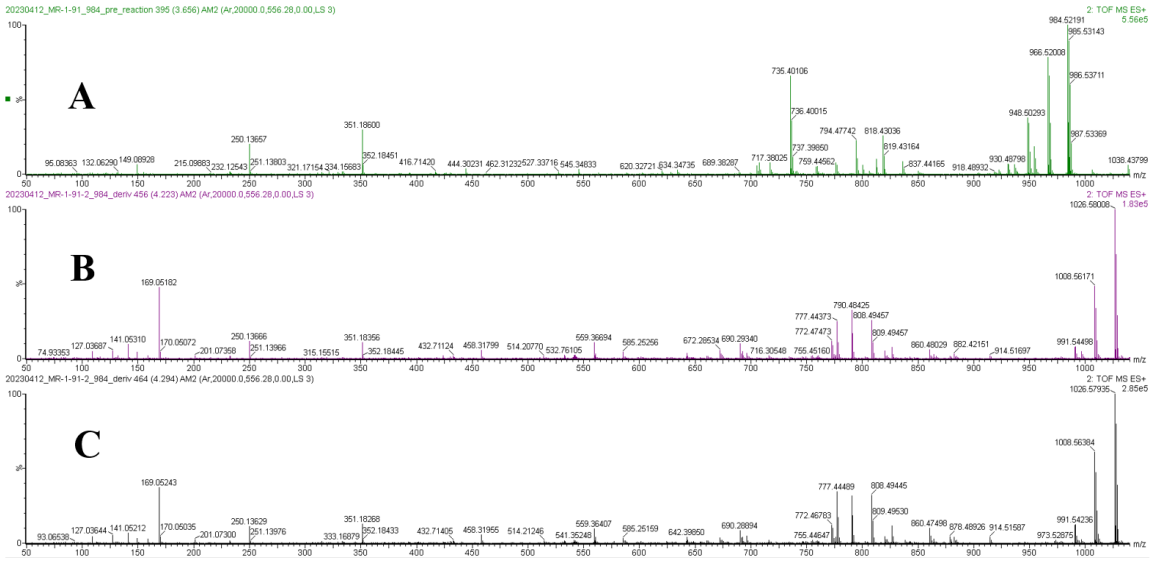


Figure A.22 Comparative MS Spectra of Megapolipeptin B (5.3) and Methyl Ester Derivatives.
(**A**) megapolipeptin B (**2**) (**B**) and (**C**) the two megapolipeptin B methyl ester derivatives (**B**: **5.6**, t_R 4.21 min; **C**: **5.7**, t_R 4.29 min).

Table A.1 **Comparison of Ground-Truth Replicate Values Versus Predicted Abundance Values from MultiplexMS-Q in the 5 x 5 "One-Compound-One-Well" Experiment.**

standard	r.t.	m/z	replicate abundance	R1	R2	R3	std dev	predicted values
Lincomycin	1.58	126.1299	13952.24	13623.06	13098.98	15134.68	863.05	13380.65
	1.58	316.2002	5449.04	5776.21	5479.12	5091.79	280.22	4620.42
	1.58	358.2112	48777.54	51915.12	48756.72	45660.77	2553.37	40796.50
	1.58	389.2105	15207.86	15971.47	15491.59	14160.53	766.05	13911.41
	1.58	407.2246	1343854.05	1415814.81	1349968.36	1265778.99	61404.27	1220320.94
	1.58	812.4229	3502.72	3701.41	3132.97	3673.78	261.70	3327.53
Doxycycline	2.27	188.0874	13248.42	13640.33	14573.75	11531.17	1272.67	10502.70
	2.27	218.0639	8534.60	9297.29	9299.91	7006.61	1080.46	7105.56
	2.27	252.0438	3594.39	4236.66	3756.86	2789.65	601.81	3458.95
	2.27	267.0668	5167.72	5889.67	5371.94	4241.56	688.16	5147.15
	2.27	284.0696	4426.66	4735.59	4616.66	3927.72	356.13	4426.54
	2.27	292.0752	3171.27	3332.02	3281.60	2900.20	192.78	3681.20
	2.27	338.0817	4623.47	4547.66	4899.81	4422.93	201.93	4060.14
	2.27	392.1129	5600.63	5383.47	5951.71	5466.72	250.56	6258.83
	2.27	428.1360	834145.83	789558.64	906869.64	806009.21	51860.19	810192.73
	2.27	445.4280	84802.60	78853.46	93792.81	81761.54	6466.95	71794.02
ceftazidime	1.54	274.5600	16804.00	20352.01	16402.68	13657.29	2747.80	17738.75
	1.54	354.0331	4005.81	4941.30	3840.48	3235.67	706.06	4197.06
	1.54	396.0813	46448.06	55713.31	44709.91	38920.94	6964.76	51401.07
	1.54	423.0674	5880.71	6962.50	5731.48	4948.14	829.10	6276.86
	1.54	440.0701	18715.81	22559.50	18098.50	15489.43	2919.16	20308.85
	1.54	468.0658	68759.48	81304.82	63606.41	61367.20	8917.88	75124.91
	1.54	524.1180	104225.78	122316.34	93756.56	96604.45	12844.68	112147.40
midecamycin	2.94	613.3416	8053.81	7864.98	8261.58	8034.88	162.46	0.00
	2.94	813.4473	24929.59	23045.62	27655.59	24087.55	1973.95	17063.85

	2.94	814.8166	129831.82	129485.12	124756.36	135254.00	4292.65	88404.47
	2.96	815.4598	1740122.38	1714416.55	1739296.54	1766654.07	21333.87	1219800.22
florfenicol	2.54	127.0567	3990.34	3726.12	3702.46	4542.45	390.52	3723.84
	2.54	169.0544	6249.32	4983.16	5841.72	7923.08	1234.34	4766.99
	2.54	205.0309	2981.08	2854.32	2893.81	3195.09	152.19	2902.92
	2.54	241.0075	67943.81	67027.70	65602.46	71201.27	2375.73	68939.25
	2.54	243.0044	45786.53	45118.32	44460.63	47780.65	1435.39	45724.66
	2.54	319.9916	5341.02	5312.16	5228.15	5482.75	105.92	5239.12
	2.54	321.9886	4233.59	4231.57	4167.48	4301.73	54.83	4127.45
	2.54	357.0020	174652.63	174970.30	168544.41	180443.20	4862.85	171534.15
	2.54	358.9993	119077.13	119643.65	114667.01	122920.72	3393.29	117438.12
nystatin	2.84	163.0861	6614.63	6222.50	6552.44	7068.95	348.35	6452.38
	2.84	164.0637	2939.00	2868.55	2848.62	3099.83	114.01	3660.91
	2.84	203.1443	2208.04	2187.99	2144.25	2291.88	61.91	0.00
	2.84	219.1762	77160.22	76981.11	73617.96	80881.59	2968.07	16847.28
	2.84	232.0912	2370.71	2209.45	2330.73	2571.95	150.66	2606.84
	2.84	264.2333	3058.17	3065.72	2869.23	3239.55	151.28	0.00
	2.84	292.1556	2223.35	2103.81	2177.91	2388.32	120.51	0.00
	2.84	297.1364	2160.46	2044.62	2171.29	2265.45	90.48	0.00
	2.84	336.1822	5267.62	5085.35	5110.82	5606.69	239.99	4674.02
	2.84	354.6952	2973.18	2864.22	2885.05	3170.27	139.62	2386.21
	2.84	356.1470	2847.73	2764.97	2768.32	3009.89	114.67	2523.62
	2.84	363.1992	11085.76	10691.01	10775.15	11791.11	499.94	8672.58
	2.84	363.7013	4108.90	3927.72	3952.23	4446.75	239.10	3262.37
	2.84	422.7590	7649.76	7334.29	7362.99	8252.01	426.01	6904.87
	2.84	446.2503	3033.32	2925.61	2992.12	3182.24	108.74	2798.86
	2.84	654.3529	4379.15	4380.58	4255.95	4500.91	100.01	3890.47
	2.84	690.3752	22809.86	22476.07	22246.70	23706.82	641.12	23533.41

	2.84	709.3921	12815.43	12555.04	12769.20	13122.04	233.77	12313.81
	2.84	744.4062	21817.52	21169.83	21400.69	22882.03	758.60	22766.94
	2.84	907.4886	179160.41	176586.73	176309.04	184585.45	3837.76	139255.21
	2.84	925.4981	1193625.73	1201245.46	1161587.18	1218044.54	23670.01	980445.48
	2.84	926.6308	9057.98	8843.31	8641.27	9689.34	454.00	7775.96
amoxicillin	1.16	114.0030	55454.10	51580.29	54410.45	60371.55	3664.10	58710.46
	1.16	114.0371	2477.68	2441.62	2359.67	2631.74	113.96	2355.50
	1.16	137.0076	2833.40	2694.98	2800.95	3004.27	128.34	2855.84
	1.16	139.0653	2432.72	2382.14	2297.26	2618.77	136.04	2471.76
	1.16	165.0669	3135.55	2980.92	3077.41	3348.32	155.52	3251.53
	1.16	180.0493	6360.05	6045.91	6157.64	6876.59	368.09	6530.65
	1.16	188.0835	24059.53	22504.13	23719.48	25954.99	1429.18	24808.05
	1.16	193.0622	2524.50	2442.44	2473.31	2657.75	95.06	2701.10
	1.16	193.1138	3116.78	3000.32	2982.02	3368.01	177.80	3197.64
	1.16	234.0234	12745.73	12127.57	12522.91	13586.70	616.17	13092.23
	1.16	255.0443	3917.22	3702.06	3856.37	4193.22	205.08	3965.46
	1.16	277.1011	6146.47	5765.75	6019.31	6654.34	373.74	6371.60
	1.16	292.1115	6674.88	6313.58	6573.21	7137.85	344.10	6953.17
	1.16	320.0842	36635.47	34418.15	36058.72	39429.54	2086.14	37182.34
	1.16	349.0870	387623.35	357976.41	389534.42	415359.24	23465.38	395581.27
	1.16	365.1057	36818.89	34226.13	36764.30	39466.24	2139.61	36906.80
	1.16	731.2157	88657.69	83442.70	85273.65	97256.71	6126.20	91201.28
ampicillin	1.61	106.0675	67055.86	68793.69	66161.08	66212.81	1229.02	65365.09
	1.61	118.0701	2665.76	2632.78	2638.03	2726.46	42.98	2608.32
	1.61	145.0550	2955.75	2852.68	2994.67	3019.90	73.60	2775.70
	1.61	160.0446	82176.29	83225.73	81706.43	81596.70	743.42	81318.19
	1.61	191.0599	84081.91	85773.39	82906.05	83566.29	1226.05	80114.05
	1.61	191.0838	14083.88	14335.44	13961.54	13954.65	177.90	13274.82

	1.61	211.0730	4341.64	4418.58	4260.21	4346.13	64.73	3986.36
	1.61	238.1169	2258.84	2287.89	2273.70	2214.93	31.58	2132.39
	1.61	266.1119	3639.66	3700.72	3584.69	3633.57	47.56	3450.37
	1.61	322.0996	11001.52	11234.07	10942.14	10828.36	170.87	10299.34
	1.61	322.1222	2958.62	3025.87	2965.92	2884.07	58.12	3123.22
	1.61	333.0919	31722.11	32549.06	30986.63	31630.63	641.13	29489.43
	1.61	349.1116	606065.91	620908.10	602157.24	595132.40	10879.80	590492.74
	1.61	350.3628	8786.67	9280.49	8383.84	8695.70	371.67	7258.89
amphotericin B	2.97	163.0858	2620.98	3540.00	2042.88	2280.08	657.02	3085.69
	2.97	353.1865	4508.85	5899.25	3658.59	3968.71	991.28	4896.05
	2.97	363.1986	3531.71	4509.61	2871.98	3213.53	705.40	0.00
	2.97	451.7310	10745.54	14226.01	7697.57	10313.04	2682.71	15747.84
	2.97	452.2330	5988.53	7674.32	4472.48	5818.79	1312.64	8018.99
	2.97	461.7437	6623.98	8669.74	4764.49	6437.72	1599.74	9765.51
	2.97	462.2457	3056.09	3911.85	2223.10	3033.34	689.62	4360.63
	2.97	688.3578	6015.01	7753.24	5078.37	5213.44	1230.35	5729.97
	2.97	724.3789	11147.37	14250.69	9452.34	9739.08	2197.50	10498.80
	2.97	760.4013	59043.92	76771.60	50184.46	50175.71	12535.36	48658.43
	2.97	905.4734	270815.92	350320.66	232255.55	229871.54	56226.77	178221.04
	2.97	921.4716	13464.81	18221.09	10076.09	12097.25	3462.94	17277.98
	2.97	932.4881	4282.69	6225.74	2863.32	3759.02	1421.77	4947.74
	2.97	957.4996	15442.95	21144.71	10628.23	14555.91	4338.91	17343.81
novobiocin	4.37	96.0384	3017.40	3199.29	2638.98	3213.92	267.65	0.00
	4.37	218.1038	23740.13	25172.40	21769.72	24278.26	1440.31	18813.83
	4.37	613.2383	655640.90	711775.47	580771.73	674375.52	55098.30	356220.00
	4.37	1225.4658	242831.28	255399.37	230701.30	242393.16	10087.70	180266.13
	4.37	1225.9623	16923.68	18824.43	14211.97	17734.64	1968.41	13955.14
chloramphenicol	2.65	142.0801	4049.82	3488.01	5154.95	3506.51	781.48	0.00

	2.65	248.0363	3481.45	3179.89	4042.92	3221.53	397.38	0.00
	2.65	275.0007	98207.28	88675.38	112265.55	93680.92	10148.56	56242.65
	2.65	276.9974	61511.34	54951.47	72071.63	57510.92	7540.00	34808.50
	2.65	278.9944	7622.76	6810.67	9014.79	7042.81	988.87	0.00
	2.65	306.0006	61084.04	56346.42	65791.28	61114.44	3855.91	35531.48
	2.65	322.0138	92716.89	86880.72	96928.58	94341.36	4259.82	55702.29
azithromycin	2.05	83.0515	17047.54	20983.48	15609.52	14549.63	2816.56	18831.82
	2.05	115.0773	4509.27	5869.13	3917.26	3741.43	964.24	4679.14
	2.05	127.0774	3004.79	4566.99	2306.33	2141.05	1106.70	2763.04
	2.05	296.2155	6486.85	8286.40	5916.32	5257.82	1300.56	7396.20
	2.05	375.5147	4584.04	5915.03	4052.88	3784.21	947.52	4616.81
	2.06	375.2625	450805.08	559610.02	402601.67	390203.55	77103.02	459347.09
	2.06	375.7646	236007.43	341793.94	186558.02	179670.33	74855.19	210794.95
	2.05	590.4142	267398.46	330022.89	243285.08	228887.42	44670.55	287526.19
cefadroxil	1.29	157.0218	22677.15	21896.23	23662.30	22472.91	735.32	21635.44
	1.29	179.0416	3544.72	3323.73	3636.10	3674.33	157.04	3708.64
	1.29	183.0391	2652.28	2499.08	2767.06	2690.69	112.73	2612.51
	1.29	194.0284	2226.96	2132.64	2307.90	2240.33	72.17	0.00
	1.29	207.0374	129219.02	125930.77	134956.89	126769.40	4071.71	96388.76
	1.29	253.0289	5264.24	5032.39	5505.49	5254.83	193.26	5170.89
	1.29	256.0617	11768.14	11298.80	12190.95	11814.65	365.70	11472.39
	1.29	275.0855	2304.77	2210.49	2458.29	2245.52	109.50	2327.87
	1.29	318.0683	12292.39	11761.90	12969.07	12146.19	503.55	11687.75
	1.29	347.0711	37868.83	36688.44	39576.90	37341.15	1236.83	34955.57
	1.29	363.0900	17190.91	16589.82	17934.11	17048.79	557.93	15665.66
penicillin G	2.92	153.0834	268993.31	263908.30	271755.01	271316.63	3600.10	288206.86
	2.92	208.0986	6544.34	10545.96	5819.89	3267.17	3015.38	4588.04
	2.92	217.0657	5092.17	7901.21	4575.00	2800.31	2114.30	3639.11

	2.92	269.1519	3589.40	3756.40	3542.17	3469.62	121.75	3731.66
	2.92	334.0997	113795.60	135852.63	109845.40	95688.78	16633.04	103901.41
	2.92	367.1340	353609.39	528898.01	330658.98	201271.17	134734.00	258886.07
deferoxamine	1.63	101.0861	4125.67	4284.88	4416.70	3675.43	322.88	4784.13
	1.63	144.1041	5238.43	5516.28	5543.28	4655.72	412.18	6221.89
	1.63	160.1242	2765.90	2849.86	2973.33	2474.51	212.12	3308.88
	1.63	165.1038	2853.32	2918.61	3119.89	2521.45	248.64	3359.94
	1.63	201.1260	66962.46	68894.85	72264.70	59727.83	5297.42	78914.73
	1.63	260.1390	92100.75	95326.63	100473.48	80502.15	8466.33	106949.20
	1.63	281.1853	7558.71	7833.61	8321.65	6520.88	760.43	8414.49
	1.63	307.6411	43915.54	46099.29	45489.56	40157.76	2668.78	42986.97
	1.63	308.1423	12518.22	13055.96	12975.60	11523.09	704.42	12568.34
	1.63	319.2353	199548.91	205972.60	214954.37	177719.76	15865.10	236854.20
	1.63	378.2505	4966.93	5145.53	5296.68	4458.58	364.72	5773.54
	1.63	443.2493	2285.59	2334.12	2443.55	2079.10	152.69	0.00
	1.63	560.3535	503880.74	509694.83	550568.72	451378.67	40702.33	586415.87
	1.63	561.6743	4757.04	4927.68	5541.57	3801.88	720.40	4201.22
cyclosporine A	5.77	1075.7402	14327.28	12796.23	15443.87	14741.74	1119.92	15399.58
	5.77	1201.8361	25561007.45	23440212.25	26520729.35	26722080.74	1501879.89	22173804.96
	5.77	1203.2921	159288.88	138838.15	170775.76	168252.74	14497.49	164820.82
cloxacillin	3.48	468.0998	162154.31	213985.50	89494.52	182982.93	52914.23	165752.84
	3.48	469.0901	64658.80	87725.56	33015.75	73235.08	23143.83	73445.68
cycloheximide	2.72	114.1078	2901.91	2723.88	2862.62	3119.23	163.77	4411.93
	2.72	141.0715	5730.39	5096.47	6270.87	5823.84	483.98	7907.37
	2.72	145.0672	6914.15	6416.01	7159.99	7166.46	352.25	9323.98
	2.72	148.0883	2693.17	2419.15	2956.31	2704.04	219.43	4003.84
	2.72	157.1019	4570.23	3998.69	5201.00	4510.99	492.62	5977.12
	2.72	159.1184	7222.59	6530.08	8097.63	7040.05	652.84	8541.30

	2.72	162.1094	5417.51	4927.41	5946.42	5378.71	416.91	6801.59
	2.72	186.1060	11152.77	10162.66	12201.38	11094.27	833.33	14937.62
	2.72	188.0846	3329.24	3343.65	3252.82	3391.23	57.42	4002.16
	2.72	190.1609	30229.73	27298.91	33781.45	29608.81	2682.66	35486.44
	2.72	200.1213	13831.91	12556.53	15105.55	13833.64	1040.63	18226.52
	2.72	236.1446	24997.60	23171.05	26676.95	25144.79	1435.06	28672.13
	2.72	246.1507	227102.04	206001.78	251590.42	223713.93	18765.05	257101.53
	2.72	265.1510	127283.44	115257.65	140569.06	126023.62	10371.67	140495.53
	2.72	282.1716	393923.04	362168.61	425048.73	394551.80	25674.55	421683.65
	2.72	304.1526	41468.35	43653.28	34338.01	46413.77	5166.33	54872.84
tetracycline	1.94	139.0563	7251.26	6531.18	7331.29	7891.32	558.15	8276.75
	1.94	152.0640	20784.06	18836.26	20700.41	22815.49	1625.59	23819.47
	1.94	154.0528	2573.74	2333.57	2823.31	2564.34	200.04	2837.61
	1.94	158.0980	9424.52	8298.25	11681.75	8293.55	1596.10	9524.17
	1.94	164.0642	10878.61	9775.86	11370.88	11489.10	781.26	12178.19
	1.94	176.0637	7065.26	6429.39	7454.31	7312.07	453.36	7770.32
	1.94	180.0590	8329.03	7532.48	9189.18	8265.43	677.84	8924.98
	1.94	195.0830	3127.29	2588.53	4273.60	2519.75	811.05	2962.06
	1.94	203.0749	33093.97	27168.72	46173.34	25939.84	9262.11	30998.96
	1.94	254.0577	3388.87	3064.66	4101.31	3000.64	504.45	3321.95
	1.94	259.0627	5075.80	4709.39	6278.34	4239.66	871.68	5097.47
	1.94	280.0750	2839.21	2489.99	3289.76	2737.88	334.27	3109.21
	1.94	308.0690	2817.96	2570.28	2973.70	2909.89	177.06	3157.96
	1.94	382.0728	5230.67	4957.15	5141.22	5593.65	267.44	5973.07
	1.94	444.4238	35582.57	32334.21	34855.45	39558.06	2993.61	38988.86
	1.94	445.2610	20683.38	24589.19	13082.87	24378.08	5375.06	22929.46
clarithromycin	2.92	558.3628	7789.96	8419.36	6281.49	8669.04	1071.51	7876.03
	2.92	748.8341	45709.83	51292.07	32083.94	53753.49	9687.22	43571.48

erythromycin	2.58	158.1190	2666.92	2356.82	2749.71	2894.23	227.07	3623.14
	2.58	540.3524	12202.20	11552.87	12557.15	12496.57	459.81	14530.35
	2.58	575.3678	267012.04	254622.42	274343.98	272069.73	8809.85	196382.98
	2.58	715.4490	118647.20	114227.61	124776.60	116937.40	4473.09	146811.20
	2.58	734.8122	100364.92	94983.76	104398.87	101712.14	3959.99	118765.46
clindamycin	2.20	126.1299	10275.30	10494.48	10458.38	9873.05	284.82	9327.37
	2.20	335.1736	3029.34	2941.16	3185.48	2961.38	110.72	2636.85
	2.20	377.1846	18942.56	18398.06	20241.09	18188.53	922.17	16041.12
	2.20	379.1730	5654.65	5533.46	5994.16	5436.33	243.32	4997.16
	2.20	388.2032	6270.32	6090.01	6650.91	6070.02	269.25	5952.66
	2.20	425.3679	43053.66	38752.08	41672.14	48736.77	4191.65	36799.22
	2.20	425.4467	179173.57	172007.67	188755.46	176757.58	7047.45	153083.65
	2.20	425.6157	15770.01	15898.98	16518.80	14892.23	670.28	11898.84
	2.20	426.1913	1964001.72	1868479.19	2128331.55	1895194.41	116709.46	1244395.20
roxithromycin	2.93	83.0513	40177.47	37954.49	36916.55	45661.37	3900.79	21424.41
	2.93	119.1560	29849.09	27854.61	27503.76	34188.90	3072.05	16190.85
	2.93	308.1970	3421.87	3258.66	3248.07	3758.87	238.34	
	2.93	357.2263	5061.16	4851.66	4647.50	5684.30	448.44	
	2.93	366.2396	59042.93	56871.51	53358.54	66898.75	5737.05	29847.53
	2.93	366.7413	21046.56	20470.10	18943.94	23725.65	1994.23	10576.95
	2.93	381.2440	6578.21	6180.95	6051.27	7502.41	655.65	
	2.93	381.7455	2198.47	2057.55	2007.85	2530.01	235.31	
	2.93	419.2711	322606.92	321335.93	292795.86	353688.98	24875.75	167831.39
	2.93	419.7721	140233.77	138849.27	127119.67	154732.36	11315.26	70993.18
	2.93	514.3239	7730.23	7288.60	7239.15	8662.95	659.84	
	2.93	572.3675	153174.25	144954.24	141739.32	172829.19	13959.97	79958.34
	2.93	678.4295	411181.98	375112.20	397773.80	460659.93	36188.74	232826.64
	2.93	715.4482	5149.45	4229.49	5250.33	5968.53	713.53	

	2.93	836.5203	42516.09	43483.36	33216.06	50848.87	7230.98	20374.97
	2.93	837.8956	168509.62	135522.93	174492.55	195513.37	24853.71	86151.11
piperacillin	2.77	1035.3302	403565.44	376989.57	435456.20	398250.54	24162.96	379435.47
	2.77	1057.3088	7856.87	7289.68	8602.84	7678.08	550.80	7452.80
	2.77	143.0847	56515.18	55171.04	58251.21	56123.29	1287.65	54632.20
	2.77	159.0372	32029.75	30839.57	33165.00	32084.68	950.15	30352.71
	2.77	274.1198	3510.42	3441.59	3566.31	3523.35	51.73	3530.84
	2.77	302.1170	14455.96	13633.43	15431.15	14303.29	741.81	12963.01
	2.77	359.1355	40049.22	38293.91	42442.35	39411.41	1752.62	35450.68
	2.77	490.1655	4154.75	4036.23	4341.62	4086.39	133.72	2932.45
	2.77	518.1703	162209.03	153928.29	174520.46	158178.35	8876.72	140163.35
mupirocin	3.26	114.0497	2774.43	2674.09	2669.82	2979.38	144.93	2879.93
	3.26	148.1146	5494.66	5278.36	5253.90	5951.72	323.34	5814.38
	3.26	160.0900	2290.56	2189.38	2235.06	2447.23	112.34	2384.46
	3.26	180.1400	2678.97	2607.70	2706.09	2723.12	50.87	2942.62
	3.26	212.1086	4527.03	4456.57	4496.24	4628.29	73.41	4847.55
	3.26	226.1248	16546.15	16194.11	16717.94	16726.40	248.96	17694.36
	3.26	248.1284	4056.12	4110.18	4085.09	3973.10	59.59	4200.81
	3.26	262.1573	2784.95	2740.28	2811.40	2803.17	31.76	2959.68
	3.26	265.1461	18899.45	18702.56	19000.31	18995.48	139.24	20286.34
	3.26	273.1502	25874.24	25380.49	25980.59	26261.63	367.50	27771.71
	3.26	308.1635	207996.96	204039.46	210545.41	209406.00	2836.77	225308.14
	3.26	337.2010	3191.82	3185.83	3243.15	3146.48	39.69	3359.77
	3.26	344.1852	276776.01	268255.53	281891.49	280181.03	6065.23	301221.30
	3.26	402.2398	2352.69	2281.22	2375.26	2401.58	51.66	2508.38
	3.26	447.2748	38901.38	38107.36	39341.21	39255.56	562.54	42003.08
	3.26	482.2896	378442.37	369239.47	383381.42	382706.21	6513.27	405752.29
	3.26	501.3064	1270962.77	1215252.98	1292490.06	1305145.28	39730.13	1396364.45

Table A.2 **Comparison of Ground-Truth Replicate Values Versus Predicted Abundance Values from MultiplexMS-Q in the 10 x 10 Complex Background Experiment.**

standard	r.t.	m/z	R1	R2	R3	R4	IQR3	IQR1	quartile average	std dev	predicted values
lincomycin	1.59	407.2211	30307.02	48339.02	48946.49	53756.02	50148.87	43831.02	50347.18	2423.14	52882.53
deferoxamine	1.63	561.3597	2395.72	2188.20	1591.39	1289.13	2240.08	1515.82	1866.11	445.14	24978.95
tetracycline	1.95	445.1601	23060.77	22250.37	21162.68	25784.48	23741.70	21978.44	23064.57	1708.66	20606.66
azithromycin	2.06	375.2616	22806.34	12848.71	11627.44	12594.37	15338.12	12352.64	12356.84	526.11	15428.54
clindamycin	2.23	425.1877	54003.72	61596.72	60820.57	59510.61	61014.61	58133.89	58982.90	2969.84	55197.58
erythromycin	2.60	733.4616	215045.72	232565.07	199629.11	202280.53	219425.56	201617.67	212380.11	13030.04	206982.01
nystatin	2.86	925.5027	56871.17	139520.94	52907.69	55312.27	77533.61	54711.12	55030.37	1630.32	58180.39
roxithromycin	2.96	836.5241	99483.28	105199.52	116742.95	90696.27	108085.38	97286.53	103030.51	9453.20	92527.39
mupirocin	3.26	483.2724	35336.94	31701.94	35615.09	34587.66	35406.48	33866.23	34310.41	1552.17	31271.15
novobiocin	4.37	613.2389	23945.50	49011.32	26931.41	37715.50	40539.45	26184.94	34400.93	9868.63	34802.67

Table A.3 ^{1}H (600 MHz) and ^{13}C NMR (150 MHz) Data for Megapolipeptin A (5.2) in DMSO-d_6[a]

	position	δ_H, (*J* in Hz)	δ_C, mult
Ahpa	1		173.4, qC
	2a	2.26, dd (14.6, 2.8)	39.9, CH_2
	2b	2.08, dd (14.6, 9.8)	
	3	3.85, m	67.5, CH
	4	3.64, m	55.1, CH
	5a	3.49, m	60.4, CH_2
	5b	3.49, m	
	4-NH	7.58, d (9.0)	
	1-NH_2	7.19, s	
		6.76, s	
^{1}Thr	1		169.8, qC
	2	4.21, m	58.6, CH
	3	3.84, m	67.1, CH
	4	1.04, d (6.4)	19.6, CH_3
	2-NH	7.86, brd (7.6)	
^{2}Thr	1		169.9, qC
	2	4.32, dd (8.4)	57.6, CH
	3	3.93, m	66.9, CH
	4	0.96, d (6.3)	19.2, CH_3
	2-NH	7.93, brs	
Adhda	1		170.5, qC
	2a	2.56, m	39.3, CH_2
	2b	2.36, m	
	3	4.21, m	44.4, CH
	4a	1.51, m	41.8, CH_2
	4b	1.51, m	
	5	3.39, m	66.6, CH
	6a	1.37, m	36.8, CH_2
	6b	1.23, m	
	7a	1.25, m	29.2, CH_2
	7b	1.25, m	
	8a	1.33, m	24.5, CH_2
	8b	1.23, m	
	9a	1.91, m	32.0, CH_2
	9b	1.91, m	
	10	5.36, m	130.3, CH
	11	5.35, m	129.6, CH
	12a	1.99, brs	32.1, CH_2
	12b	1.99, brs	

	13a	1.99, brs	32.1, CH_2
	13b	1.99, brs	
	14	5.35, m	129.4, CH
	15	5.36, m	130.2, CH
	16a	1.90, m	32.1, CH_2
	16b	1.90, m	
	17a	1.23, m	24.6, CH_2
	17b	1.23, m	
	18a	1.42, m	34.8, CH_2
	18b	1.26, m	
	19	3.45, m	75.0, CH
	20	1.04, d (6.4)	20.2, CH_3
	21		159.8, qC
	3-NH	8.49, d (9.0)	
Hoha	1		174.0, qC
	2	4.16, m	73.0, CH
	3a	2.72, m	45.3, CH_2
	3b	2.72, m	
	4		206.1, qC
	5a	2.71, m	37.5, CH_2
	5b	2.66, m	
	6a	2.38, m	27.6, CH_2
	6b	2.38, m	
	7		173.7, qC
Mpo	1		202.3, qC
	2	3.39, m	33.9, CH
	3	1.01, d (7.2)	17.5, CH_3
	4	1.01, d (7.2)	17.3, CH_3

[a] All data was acquired on a 600 MHz NMR spectrometer equipped with a 5mm QCI cryoprobe.

Table A.4 ^{1}H (600 MHz) and ^{13}C NMR (150 MHz) Data for Megapolipeptin B (5.3) in DMSO-d_6[a]

	position	δ_H, (*J* in Hz)	δ_C, mult
Ahpa	1		173.4, qC
	2a	2.26, m	39.9, CH_2
	2b	2.08, dd (14.8, 9.5)	
	3	3.85, m	67.5, CH
	4	3.64, m	55.1, CH
	5a	3.48, m	60.4, CH_2
	5b	3.48, m	
	4-NH	7.62, d (8.7)	
	1-NH_2	7.20, s	
		6.75, s	
^{1}Thr	1		169.8, qC
	2	4.21, m	58.6, CH
	3	3.84, m	67.1, CH
	4	1.05, m	19.6, CH_3
	2-NH	7.95, brs	
^{2}Thr	1		169.9, qC
	2	4.31, m	57.7, CH
	3	3.93, m	66.6, CH
	4	0.95, d (6.0)	19.2, CH_3
	2-NH	8.04, brs	
Adhta	1		170.5, qC
	2a	2.56, m	39.3, CH_2
	2b	2.36, m	
	3	4.21, m	44.4, CH
	4a	1.52, m	41.7, CH_2
	4b	1.52, m	
	5	3.39, m	66.4, CH
	6a	1.42, m	36.9, CH_2
	6b	1.30, m	
	7a	2.03, m	28.2, CH_2
	7b	1.92, m	
	8	5.36, m	130.3, CH
	9	5.36, m	129.8, CH
	10a	1.99, m	32.1, CH_2
	10b	1.99, m	
	11a	1.99, m	32.1, CH_2
	11b	1.99, m	
	12	5.36, m	129.5, CH
	13	5.36, m	129.1, CH

	14a	1.99, m	32.1, CH_2
	14b	1.99, m	
	15a	1.99, m	32.1, CH_2
	15b	1.99, m	
	16	5.36, m	129.7, CH
	17	5.36, m	130.2, CH
	18a	1.91, m	32.0, CH_2
	18b	1.91, m	
	19a	1.23, m	24.5, CH_2
	19b	1.23, m	
	20a	1.42, m	34.8, CH_2
	20b	1.25, m	
	21	3.46, m	75.1, CH
	22	1.04, m	20.3, CH_3
	23		159.9, qC
	3-NH	8.49, d (8.7)	
Hoha	1		174.5, qC
	2	4.13, m	73.5, CH
	3a	2.72, m	45.7, CH_2
	3b	2.72, m	
	4		206.3, qC
	5a	2.71, m	37.6, CH_2
	5b	2.66, m	
	6a	2.37, m	27.8, CH_2
	6b	2.37, m	
	7		173.8, qC
Mpo	1		202.3, qC
	2	3.39, m	33.9, CH
	3	1.01, d (7.3)	17.5, CH_3
	4	1.01, d (7.3)	17.3, CH_3

[a] All data was acquired on a 600 MHz NMR spectrometer equipped with a 5mm QCI cryoprobe.

Table A.5 Bacterial Target Panel Strains and Culture Conditions.

Strain Name	Strain Designation	BSL	Growth Medium	Growth Condition
Gram-Positive				
Bacillus subtilis	ATCC 6051	1	TSB	37 °C
Enterococcus faecalis	ATCC 29212	2	BHI	37 °C
Enterococcus faecium	ATCC 6569	2	BHI	37 °C
Listeria ivanovii	BAA-139	1	BHI-A; HTM	37 °C; 5% CO_2
Staphylococcus aureus (Methicillin-Resistant)	BAA-44	2	TSB	37 °C
Staphylococcus aureus (Methicillin-Sensitive)	ATCC 29213	2	TSB	37 °C
Staphylococcus epidermidis	ATCC 14990	1	TSB	37 °C
Gram-Negative				
Escherichia coli	K-12 MG1655	1	NB	37 °C
Klebsiella aerogenes	ATCC 35029	1	NB	37 °C
Klebsiella pneumoniae	ATCC 700603	2	NB	37 °C
Ochrobactrum anthropi	ATCC 49687	1	TSB	37 °C
Providencia alcalifaciens	ATCC 9886	1	TSB	37 °C
Pseudomonas aeruginosa	ATCC 27853	2	TSB	37 °C
Salmonella enterica	ATCC 13311	2	NB	37 °C
Shigella sonnei	ATCC 25931	2	NB	37 °C
Vibrio cholerae	A1552 El Tor	2	TSB	37 °C
Yersinia pseudotuberculosis	ATCC 6904	2	BHI	37 °C

Table A.6 Antimicrobial Activities of Megapolipeptins A (5.2) and B (5.3).

Strain Name	Megapolipeptin A (5.2)	Megapolipeptin B (5.3)
Gram-Positive Bacteria		
Bacillus subtilis	>128 μM	>128 μM
Enterococcus faecalis	>128 μM	>128 μM
Enterococcus faecium	>128 μM	>128 μM
Listeria ivanovii	>128 μM	>128 μM
Staphylococcus aureus (Methicillin-Resistant)	>128 μM	>128 μM
Staphylococcus aureus (Methicillin-Sensitive)	>128 μM	>128 μM
Staphylococcus epidermidis	>128 μM	>128 μM
Gram-Negative Bacteria		
Escherichia coli	>128 μM	>128 μM
Klebsiella aerogenes	>128 μM	>128 μM

Klebsiella pneumoniae	>128 μM	>128 μM
Ochrobactrum anthropi	>128 μM	>128 μM
Providencia alcalifaciens	>128 μM	>128 μM
Pseudomonas aeruginosa	>128 μM	>128 μM
Salmonella enterica	>128 μM	>128 μM
Shigella sonnei	>128 μM	>128 μM
Vibrio cholerae	>128 μM	>128 μM
Yersinia pseudotuberculosis	>128 μM	>128 μM

Table A.7 Fungal Target Panel Strains and Culture Conditions.

Strain Name	Strain Designation	BSL	Growth Medium	Growth Condition
Aspergillus niger	DSM 737	1	PDB	22 °C
Candida albicans	ATCC 14053	2	SDB	37 °C
Purpureocillium lilacinum	DSM 846	2	PDB	22 °C
Saccharomyces cerevisiae	ATCC 9763	1	YM	37 °C

Table A.8 Antifungal Activities of Megapolipeptins A (5.2) and B (5.3).

Strain Name	Megapolipeptin A (5.2)	Megapolipeptin B (5.3)
Fungi		
Aspergillus niger	>100 μM	>100 μM
Candida albicans	>100 μM	>100 μM
Purpureocillium lilacinum	>100 μM	>100 μM
Saccharomyces cerevisiae	>100 μM	>100 μM

Appendix B.

NMR Spectra

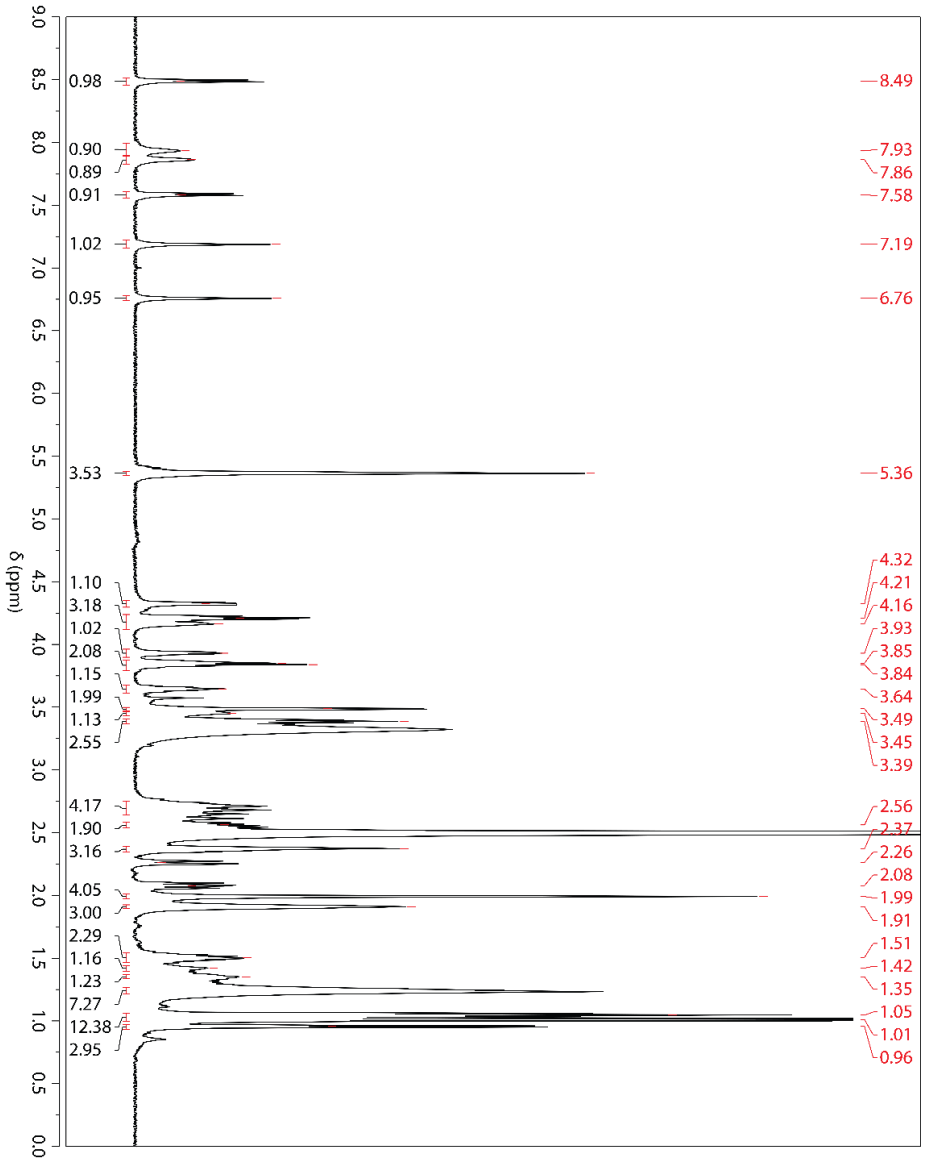


Figure B.1 **^{1}H NMR Spectrum of Megapolipeptin A (5.2) Acquired in DMSO-d_6 at 600 MHz.**

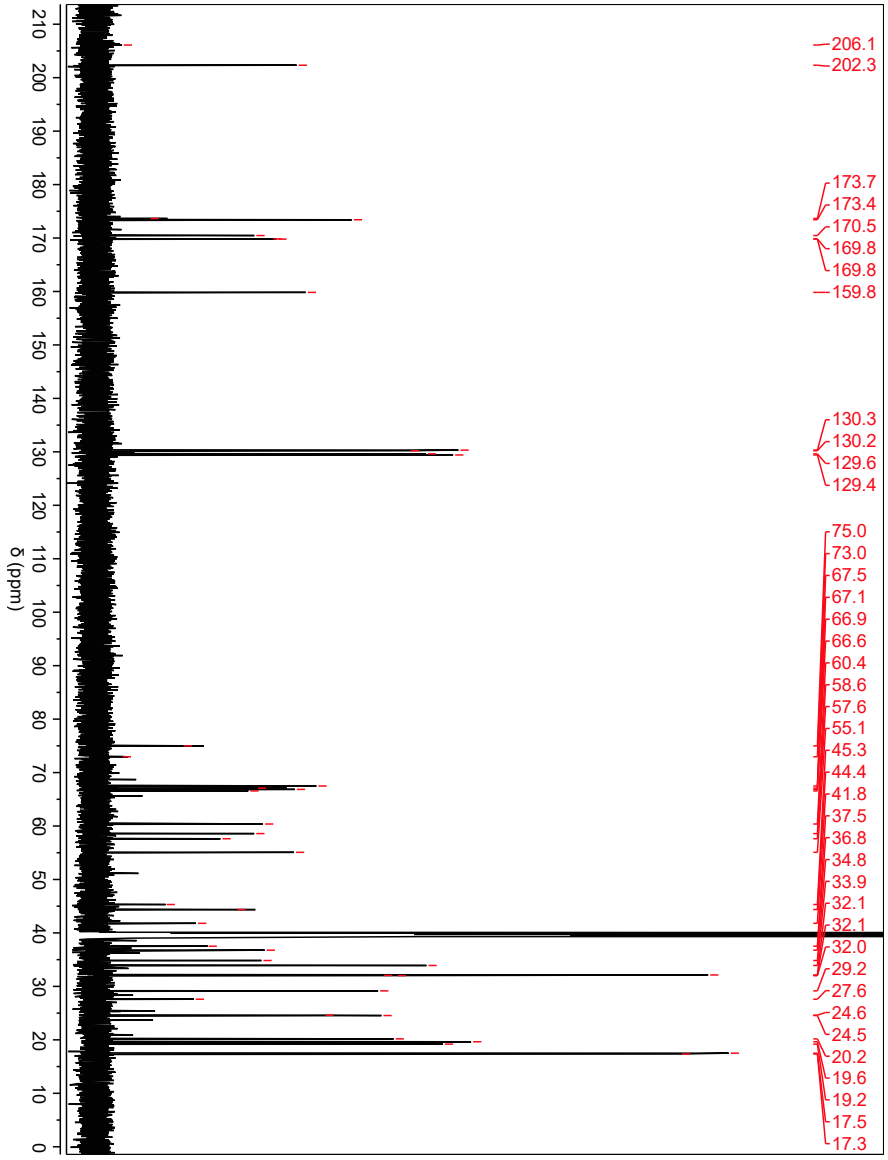


Figure B.2 **^{13}C NMR Spectrum of Megapolipeptin A (5.2) Acquired in DMSO-d_6 at 150 MHz.**

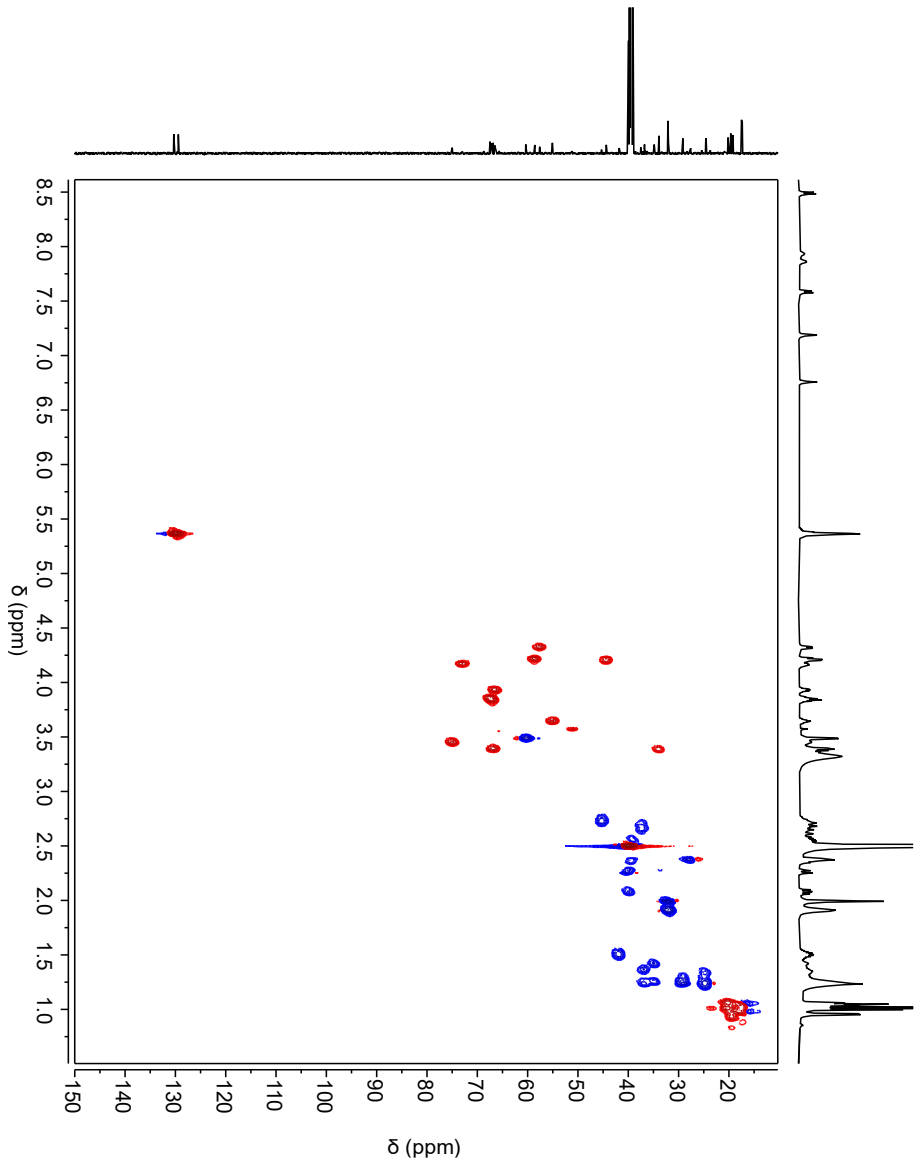


Figure B.3 **HSQC Spectrum of Megapolipeptin A (5.2) Acquired in DMSO-d_6 at 600 MHz.**

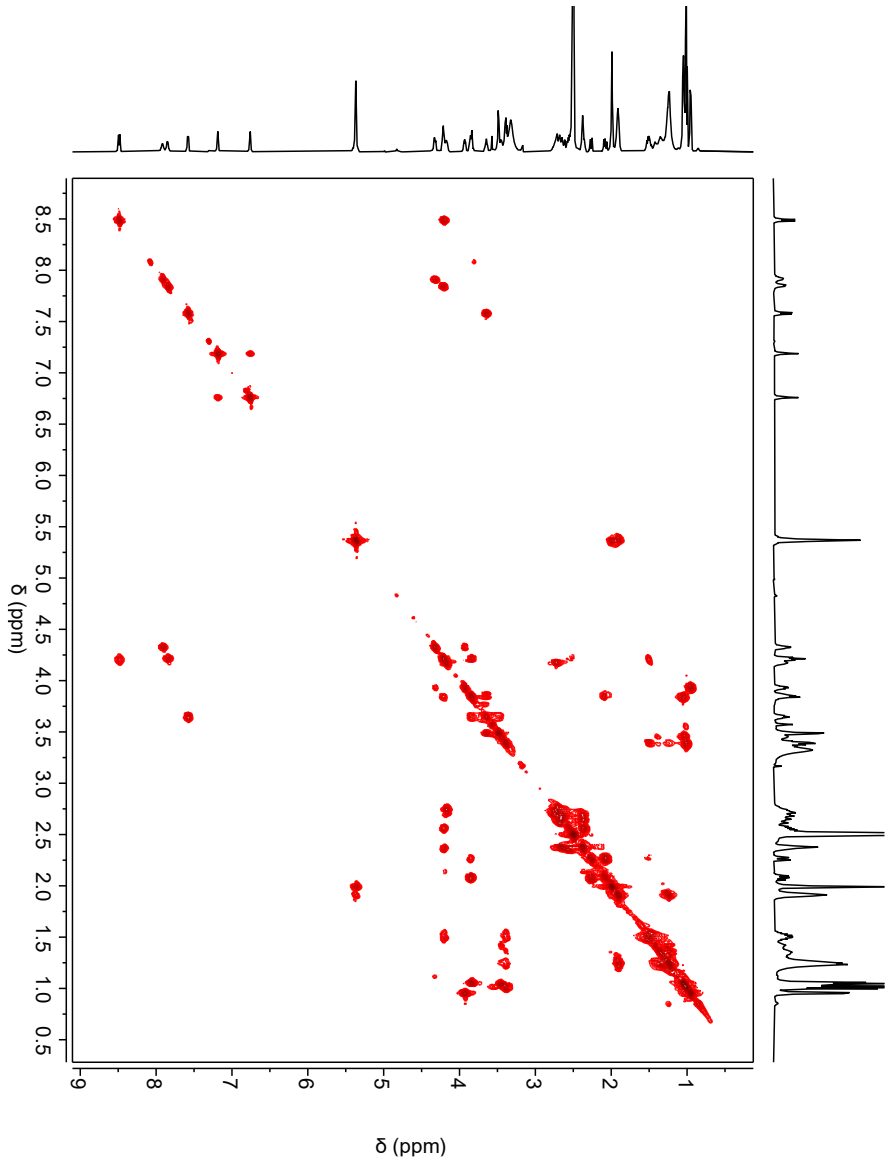


Figure B.4 **^{1}H-^{1}H COSY Spectrum of Megapolipeptin A (5.2) Acquired in DMSO-*d*$_6$ at 600 MHz.**

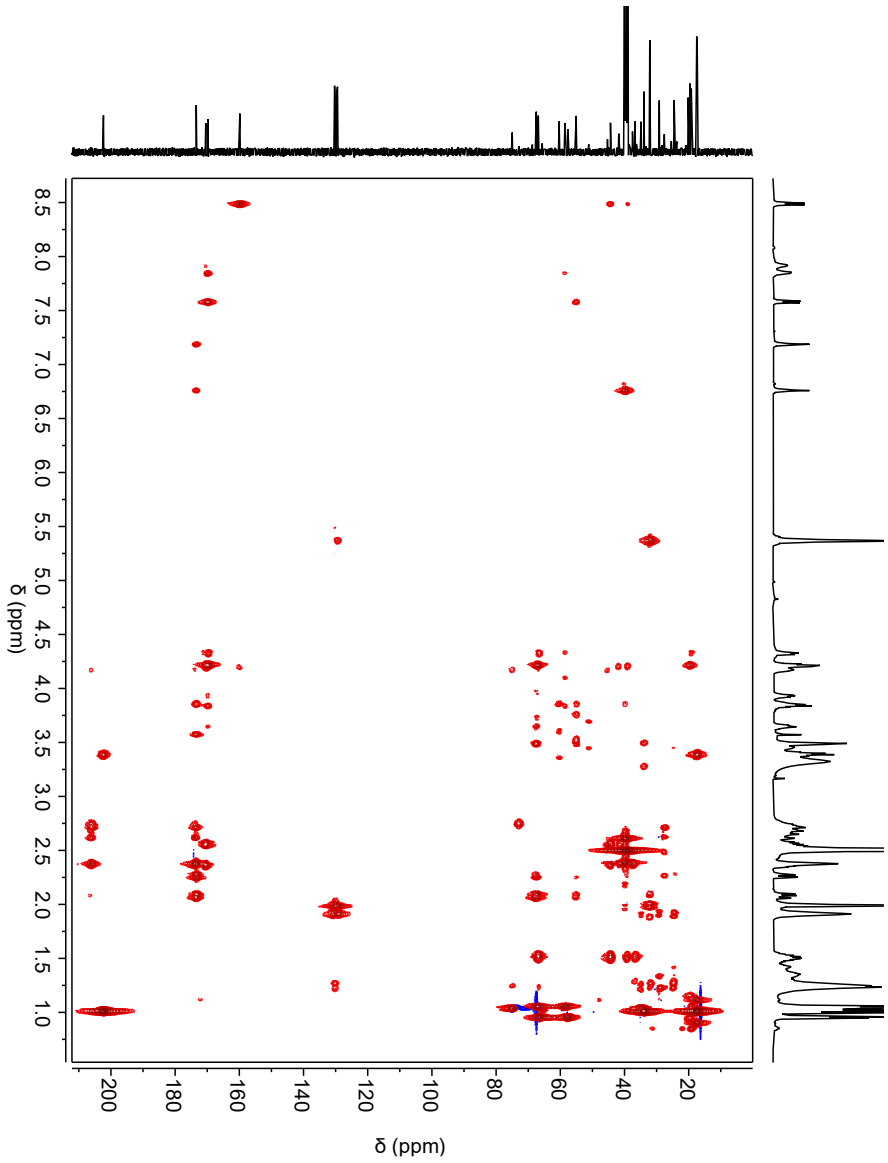


Figure B.5 **HMBC Spectrum of Megapolipeptin A (5.2) Acquired in DMSO-d_6 at 600 MHz.**

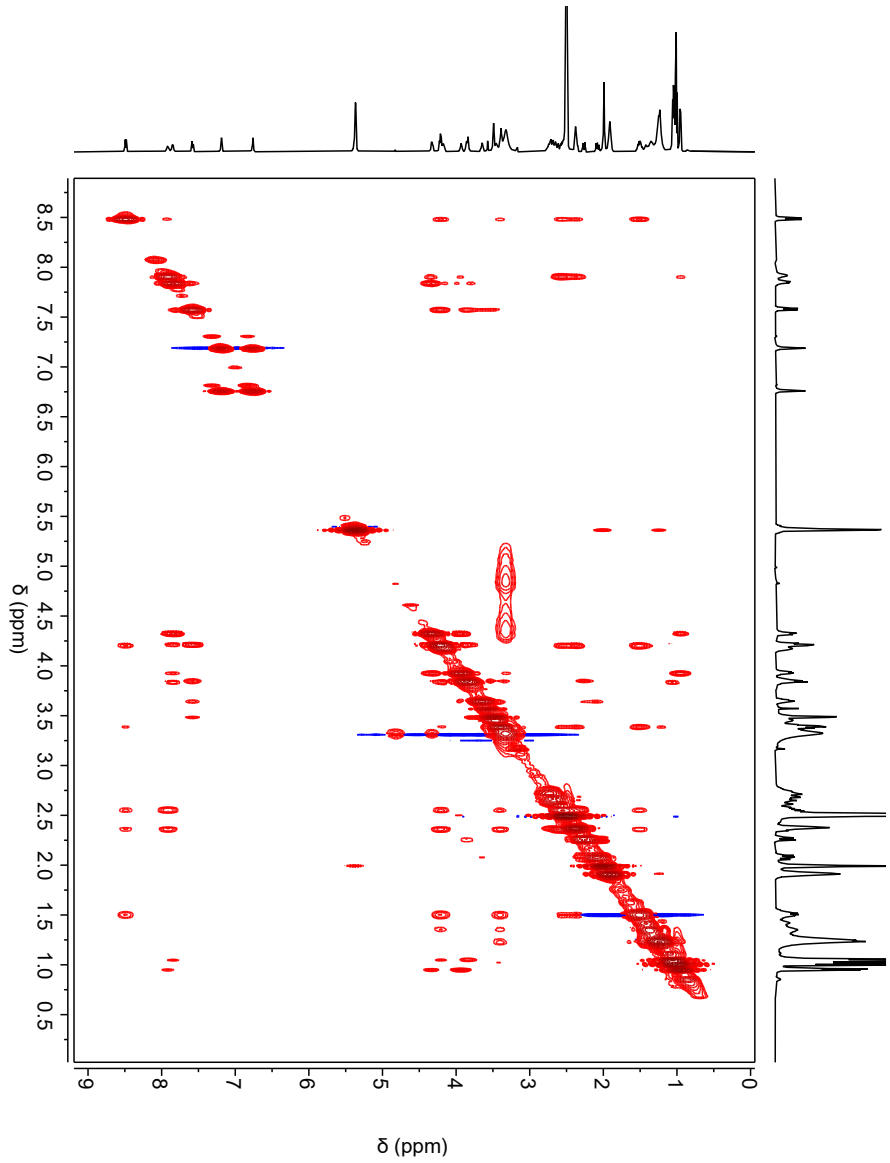


Figure B.6 **NOESY Spectrum of Megapolipeptin A (5.2) Acquired in DMSO-d_6 at 600 MHz**

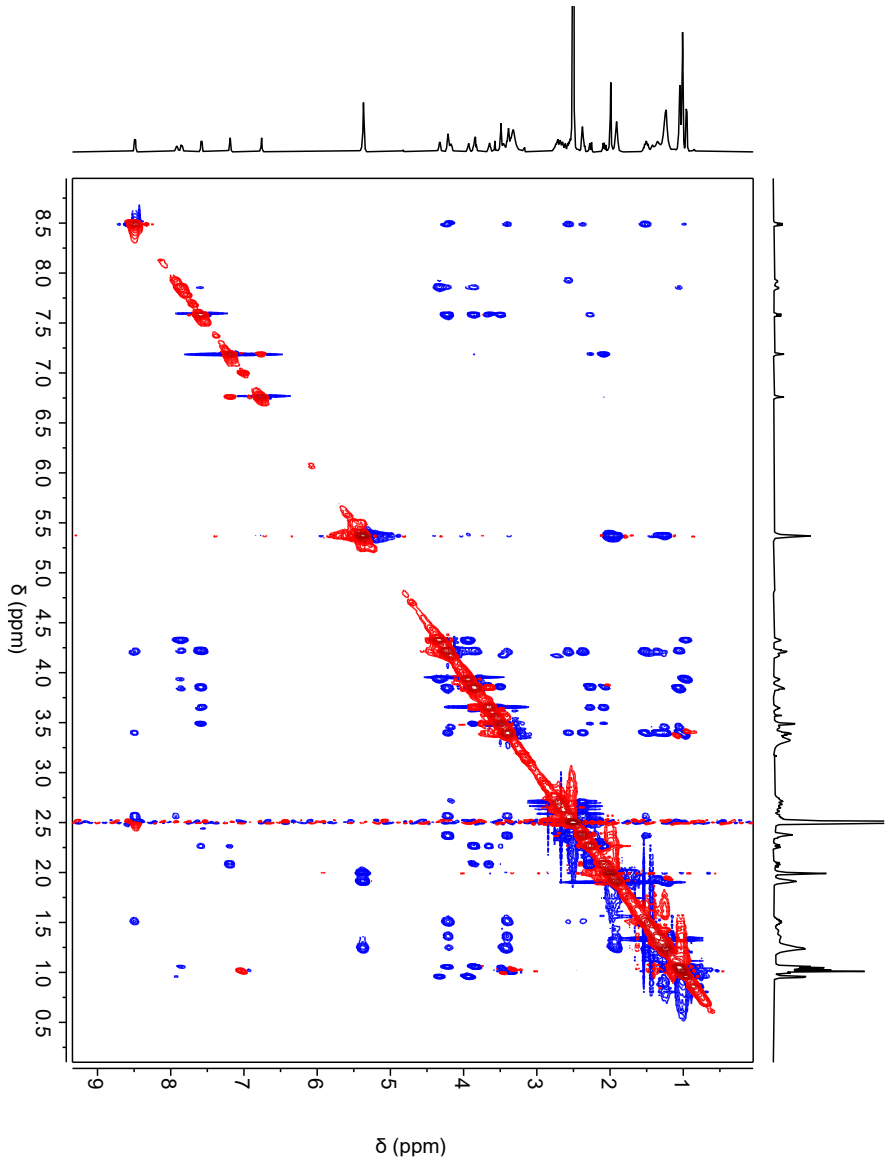


Figure B.7 **ROESY Spectrum of Megapolipeptin A (5.2) Acquired in DMSO-d_6 at 600 MHz.**

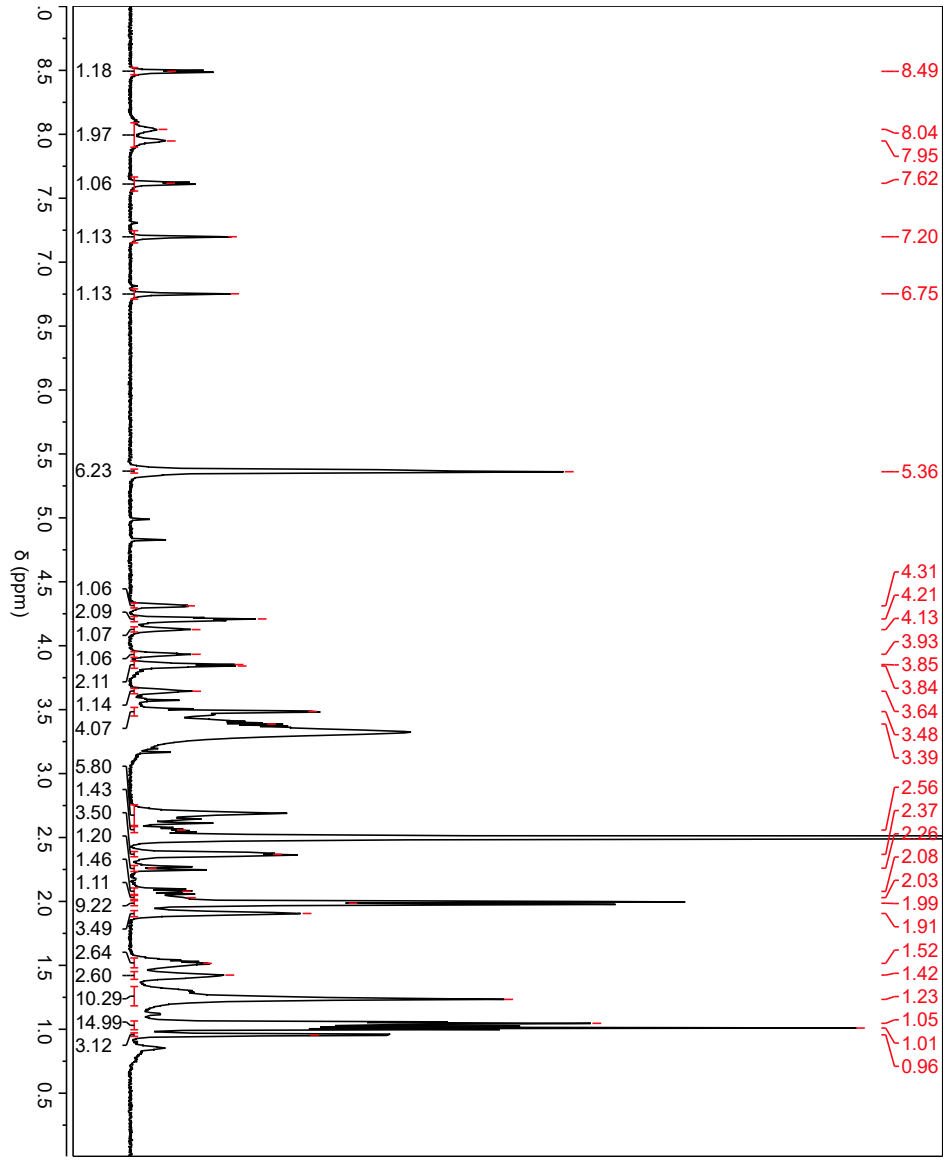


Figure B.8 **^{1}H NMR Spectrum of Megapolipeptin B (5.3) Acquired in DMSO-d_6 at 600 MHz.**

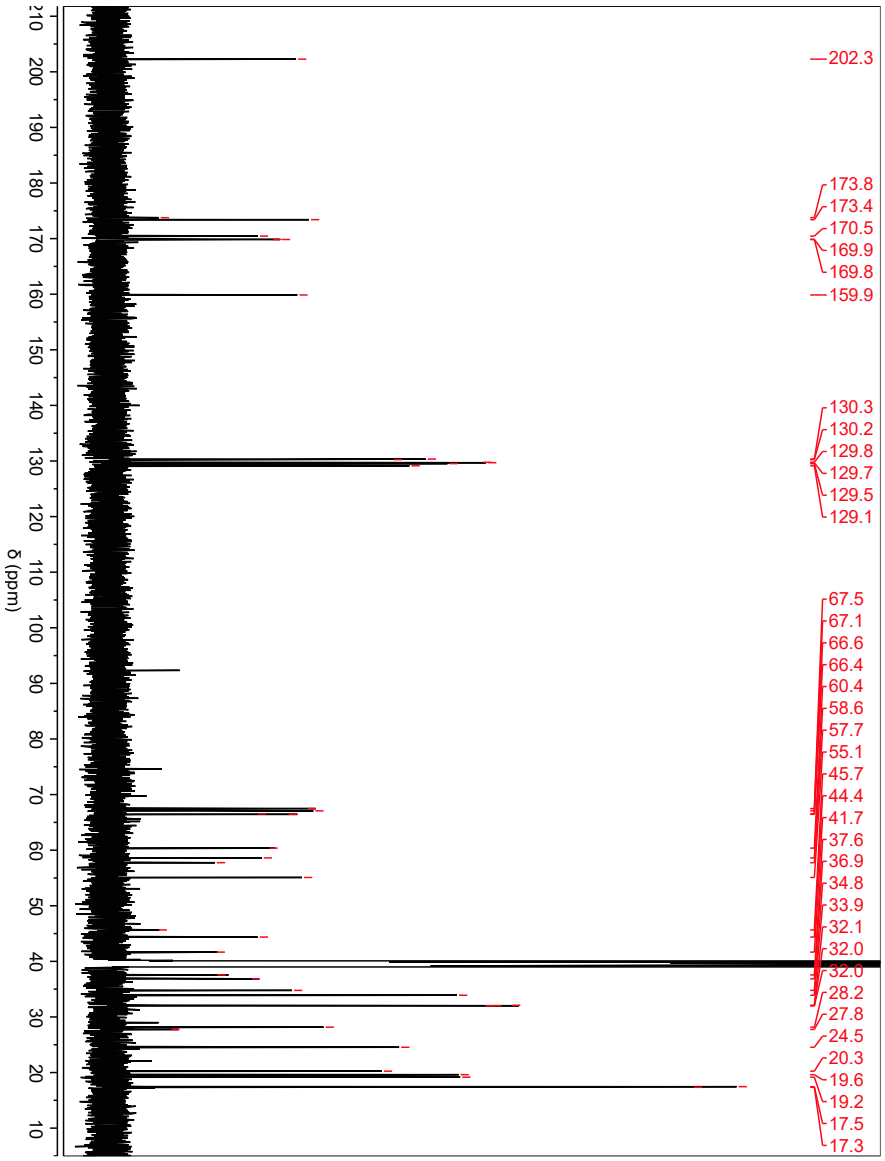


Figure B.9 **^{13}C NMR Spectrum of Megapolipeptin B (5.3) Acquired in DMSO-d_6 at 150 MHz.**

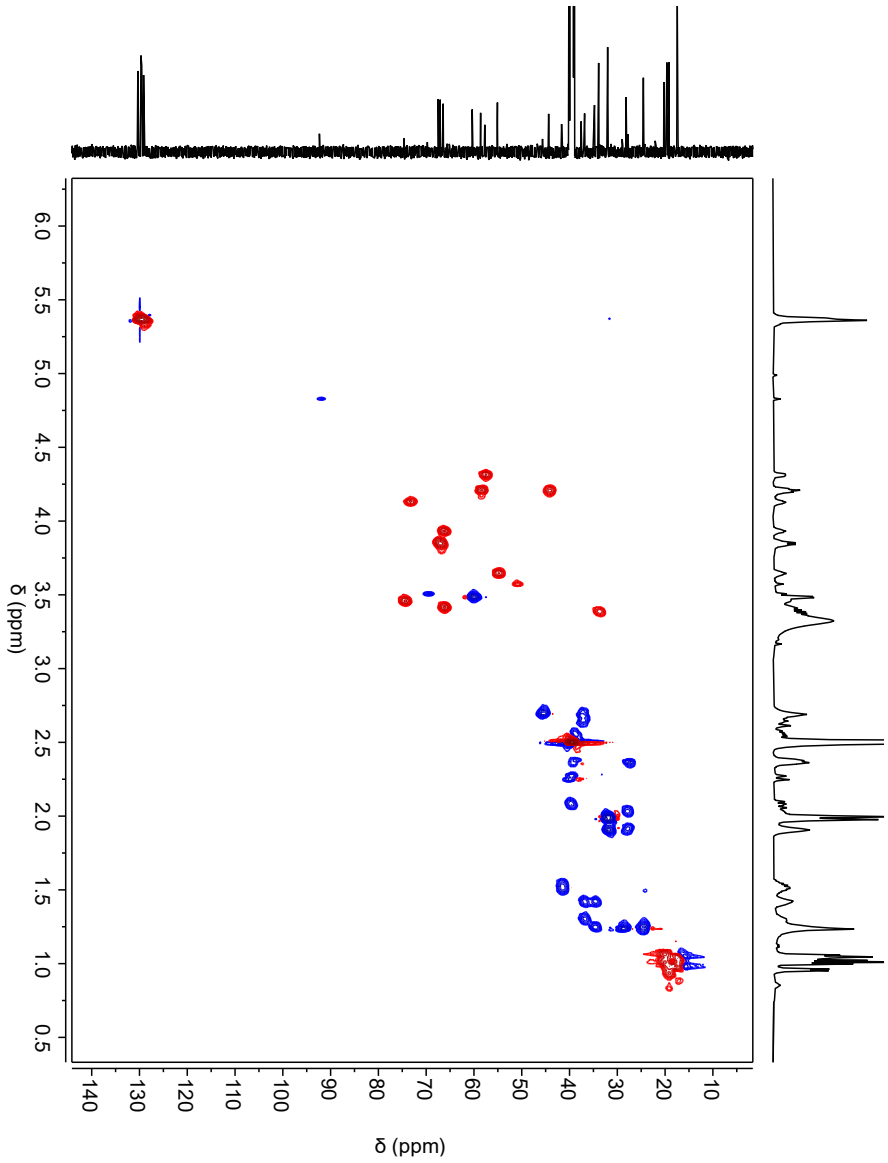


Figure B.10 HSQC Spectrum of Megapolipeptin B (5.3) Acquired in DMSO-d_6 at 600 MHz.

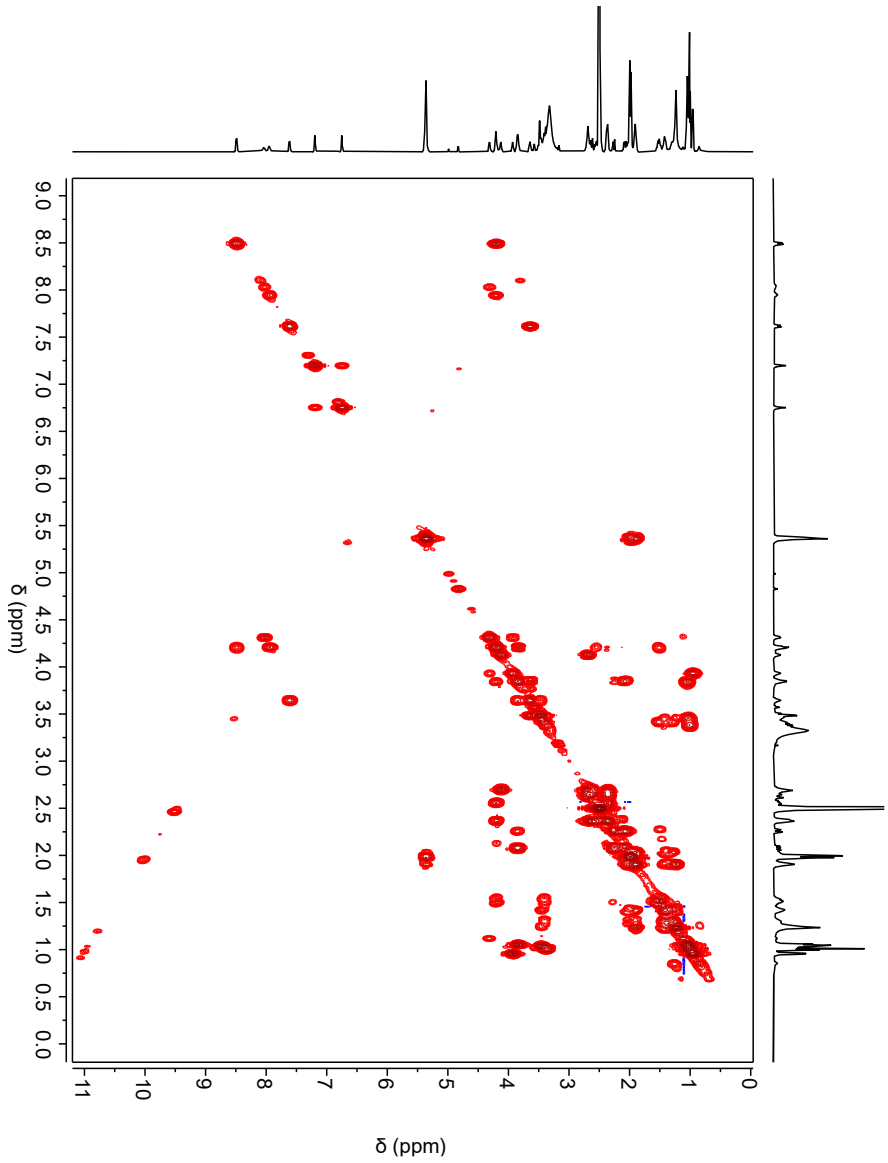


Figure B.11 **^{1}H-^{1}H COSY Spectrum of Megapolipeptin B (5.3) Acquired in DMSO-*d*$_6$ at 600 MHz.**

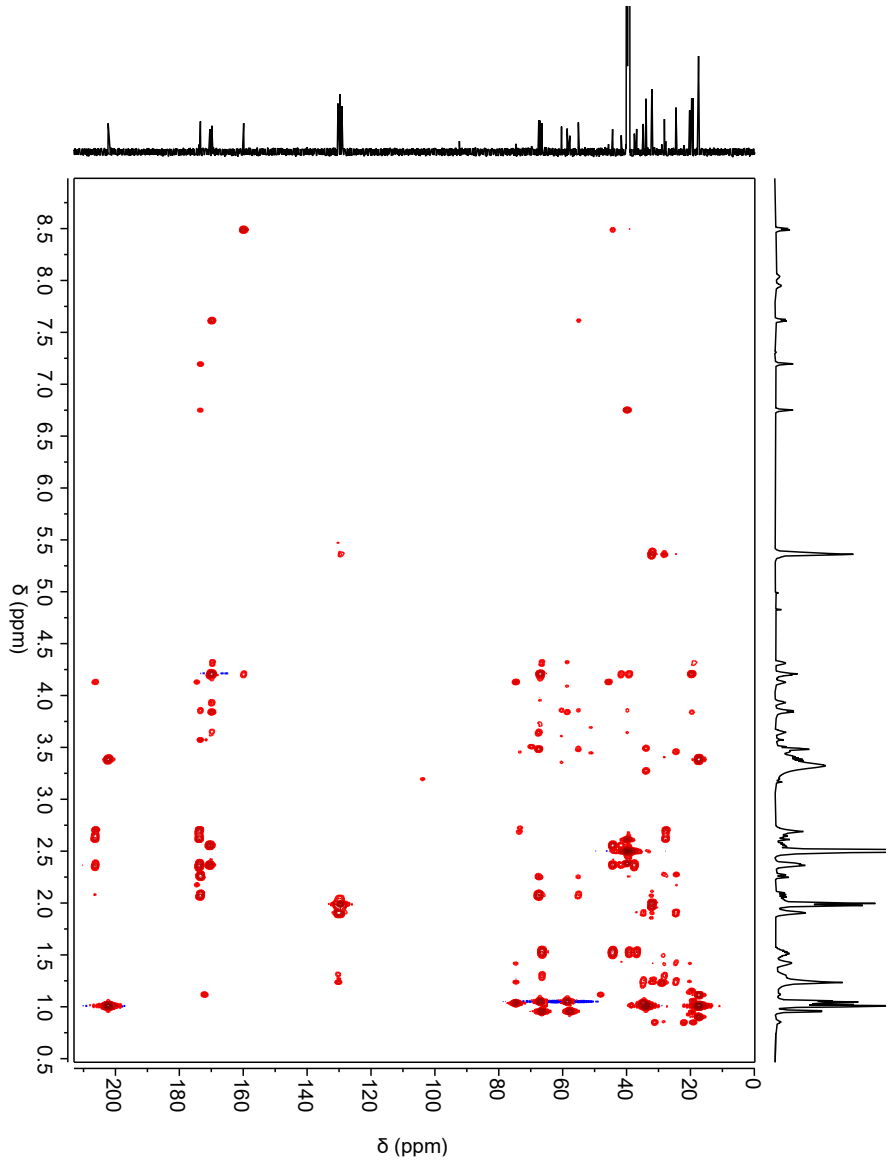


Figure B.12 HMBC Spectrum of Megapolipeptin B (5.3) Acquired in DMSO-d_6 at 600 MHz.

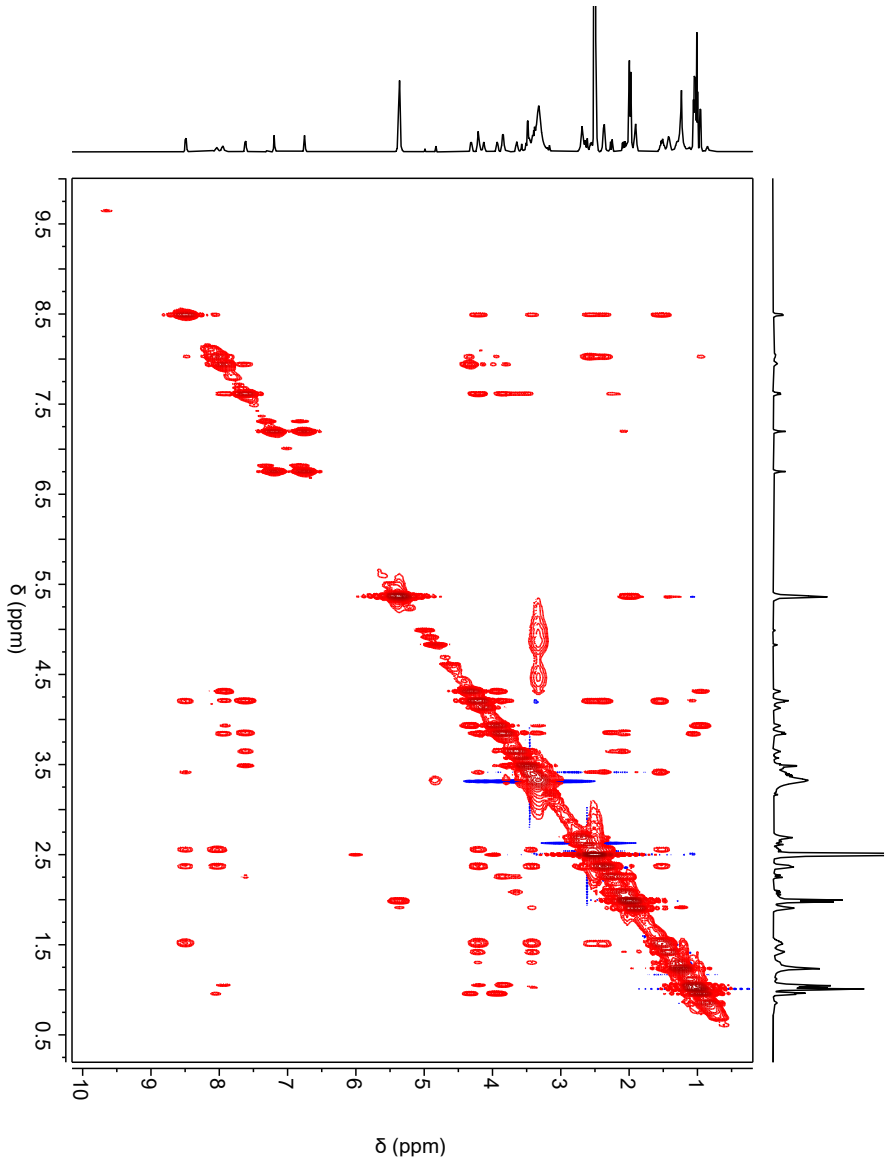


Figure B.13 NOESY Spectrum of Megapolipeptin B (5.3) Acquired in DMSO-d_6 at 600 MHz.